Establisment of DNA methylation patterns during mouse development

Rodoniki Athanasiadou

Thesis presented for the degree of Doctor of Philosophy
University of Edinburgh
2007

Declaration

I declare that this thesis was composed by myself and the research presented is my own unless otherwise stated.

Rodoniki Athanasiadou

2007

Acknowledgements

I would first like to thank Adrian, for agreeing to take me as a PhD student in the first place, for making it possible to finish the work started and for offering critical comments when I was getting too excited with my “theories”. I would also like to thank the Darwin Trust for funding my PhD. Everyone in the Bird and Stancheva labs that made the last four years I spent in the lab feel like home deserves a big thank you. I will specially mention Robert Illingworth for guidance with MAP and Miho Suzuki for lengthy and exciting discussions on CpG islands. Alastair Kerr should get the most credit for the CpG island algorithm and Dina De Sousa has offered great help with some last-minute bisulfiting. Last, but definitely not least, I would like to thank Jose de las Heras for valuable advice on microarray data processing and also for the support during these last months while I was writing this work.

Abstract

Methylation is the only known modification of DNA and in animals it mainly occurs at cytosines in a CpG context. The pattern of DNA methylation varies among organisms; some invertebrates are totally devoid of it, while others have densely methylated regions embedded in an otherwise unmethylated genome. The genome of mammals on the other hand, is very rich in DNA methylation with the exception of regions with high CpG frequency, known as CpG islands, that are often found devoid of methylation. Little is known about the factors that determine the genome-wide pattern of DNA methylation. Moreover, although there appears to be a specific developmental program for the establishment of methylation in specific genomic regions, the molecular events that lead to methylation establishment remain unknown. The establishment of methylation in the regulatory region of the murine *Oct4* gene as well as the occurrence and establishment of methylation in mouse CpG islands are investigated in this study.

The promoter of *Oct4*, which encodes an important developmental regulator, is known to gain methylation as the gene becomes silenced during early development. An *in vitro* model of murine early development has been used to recapitulate the events that lead to the gene's silencing. In accordance to other reports, detailed methylation analysis of the gene's entire upstream region and expression analysis showed that DNA methylation establishment follows the gene's downregulation. Moreover, establishment of methylation at the *Oct4* locus seems to start from the gene's proximal enhancer and then spread towards the distal enhancer and the promoter. Although the initial establishment of methylation in the distal

enhancer was not impaired in G9a -/- cells, methylation in these cells was unable to spread and accumulate. These findings demonstrate that the promoter of the gene is not the primary target for methylation as previously assumed and give rise to two possible mechanisms for DNA methylation establishment at this gene; one possibility is that methylation is actively targeted to the proximal enhancer, while the other is that the promoter and the distal enhancer are resistant to methylation, perhaps because of transcription factors bound to them. Moreover, the finding that G9a is not necessary for DNA methylation establishment but appears to have a role in methylation spreading, together with observations on the kinetics of the downregulation and the timing of methylation establishment, allowed the formation of a possible model for the role of DNA methylation in this gene's downregulation. According to this model, DNA methylation acts to accelerate the gene's downregulation ensuring its coordinated repression in the developing organism.

For the study of methylation in CpG islands, first a novel algorithm was applied for the identification of CpG islands in the mouse genome. Approximately 21,000 CpG islands were identified in the mouse genome, half of which localised at the 5' of genes, while the majority of the remaining was equally distributed in intragenic and intergenic regions. Only a very small proportion of the CpG islands localised at the 3' of genes. When the gene ontology terms related with the CpG island-associated genes where interrogated, two main gene functions emerged as being preferentially associated with CpG islands, development and cell maintenance. Then, an affinity purification method, together with microarray hybridisation was applied for the identification of methylated CpG islands from mouse brain. Approximately 18% of all CpG islands were methylated in brain, with the big majority localised at 5' and intragenic regions. When the gene ontology of the methylated CpG island-associated genes was analysed, developmental but not housekeeping genes were overrepresented in the methylated fraction. In order to further investigate the relationship of CpG islands with developmental genes, the same methodology was applied for the identification of CpG islands that become methylated after the *in vitro* induction of differentiation of ES cells. Although this approach failed to produce genome-wide data, it enforced the idea of a developmental program for CpG island methylation.

Abbreviations

A_x	absorbance at x nm
BSA	bovine serum albumin
cDNA	complementary DNA
CGBP	CpG-binding protein
ChIP	chromatin immunoprecipitation
CIP	calf intestinal phosphatase
CO	carbon monoxide
CpG	cytosine –phosphate diester– guanine
dATP	deoxyadenosine triphosphate
dCTP	deoxycytosine triphosphate
DE	distal enhancer
DEPC	diethyl pyrocarbonate
dGTP	deoxyguanosine triphosphate
DMR	differentially methylated region
DMSO	dimethylsulfoxide
DNA	deoxyribonucleic acid
DNMT	DNA methyltransferase
dNTP	deoxynucleoside triphosphate
DTT	dithiothreitol
dTTP	deoxythymidine triphosphate
E	embryonic day
EB	embryoid bodies
EC	embryonic carcinoma cells
EDTA	ethylenediamine tetraacetic acid

e.g.	(*exempla grata*) for examble
EMSA	electrophoretic mobility shift assay
ES cells	embryonic stem cells
EST	expressed sequence tag
et al.	(*et alii*) and others
FPLC	fast performance liquid chromatography
FRAP	fluorescence recovery after photobleaching
g	gram/ acceleration of gravity
GO	gene ontology
h	hours
HAT	histone acetyltransferase
HDAC	histone deacetylase
J	joules
ICF	immunodeficiency, centromeric instability and facial anomalies syndrome
ICM	inner cell mass
i.e.	(*id est*) that is
IPTG	isopropyl-β-d-thiogalactopyranoside
KS	Kolmogorov-Smirnov
l	litre
LB	Luria-Bertani
LIF	leukaemia inhibitory factor
M	molar
m	mili (10^{-3})
MAP	MBD affinity purification
MBD	methyl-CpG binding domain
min	minutes
MLL	mixed lineage leukaemia
MOPS	3-[N-morpholino]propanesulfonic acid
NMR	nuclear magnetic resonance
NOR	nucleolus organising region
nt	nucleotides
o/e	observed versus expected CpG ratio
OD_x	optical density at x nm
P	promoter
PAGE	polyacrylamide gel electrophoresis
PCR	polymerase chain reaction

PE	proximal enhancer
PGC	primordial germ cells
qPCR	real-time quantitative PCR
RA	all-trans retinoic acid/ RA-induced embryoid bodies
RARE	retinoic acid response element
RdDM	RNA-directed DNA methylation
RIP	repeat-induced point mutation
RLGS	restriction landmark genome scanning
RNA	ribonucleic acid
RNAi	RNA interference
RNase	ribonuclease
rpm	revolutions per minute
rRNA	ribosomal RNA
RT-qPCR	reverse-transcription real-time quantitative PCR
SAM	S-adenosyl-L-methionine
SDS	sodium doedecyl sulfate
sec	second
siRNA	small interfering RNA
SRA	SET- and RING- associated
T_{an}	annealing temperature
TEMED	N,N,N,N',N'- tetramethylethylenediamine
tRNA	transfer RNA
TSA	trichostatin A
U	unit
UTR	untranslated region
V	volts
v	volume
w	weight
wt	wild-type
Xa	active X chromosome
Xi	inactive X chromosome
Xic	X inactivation centre
µ	micro (10^{-6})
^{o}C	degrees centigrade

Table of contents

List of figures

List of tables

CD contents

File name	File format	File size
CpG islands	.txt (tab delimited)	1,608 Kb
CpG islands methylated in brain	.txt (tab delimited)	505 Kb
key to tables	.txt (tab delimited)	1 Kb

1. Introduction

1.1. DNA sequence in a biological context

Since the classical work of Watson and Crick in 1953 and the experiments of Marshall Nirenberg in the early 60s, the most celebrated function of DNA sequence, is containing the genetic information encoded in triplets of nucleotides. Despite the suitability of DNA's design for encoding genetic information, only approximately 1.2-21 % of the genome is dedicated to this function in higher organisms (Arabidopsis Genome Initiative 2000; Mouse Genome Sequencing Consortium 2002; Human Genome Sequencing International Consortium 2004). These numbers make it obvious that the function of DNA in an organism extends beyond being translated to protein. In fact, it is long known that the genome also encodes RNA that has either a structural, *e.g.* ribosomal RNA, or functional role, *e.g.* ribozymes. Other DNA functions include chromosome organisation, *e.g* centromeres and telomeres, and regulation of transcription, *e.g.* promoters and enhancers. However, the exact function of every genomic region is far from understood.

One way of investigating the genome function is by identifying signature nucleotide sequences. For example it is possible to test for the presence of specific transcription factor binding sites in the promoter of a gene, calculate the nucleosome positioning potential of a given sequence or identify satellite DNA sequences that comprise the mammalian centromeres. It is nevertheless still not possible to draw any certain conclusions about the function of those recognisable sequences without experimentally testing them.

There are two main reasons for the disagreement between our theoretical expectation and the actual biological role of a DNA sequence in the genome. The first reason is that, despite the extensive current proteomics studies, we are still far from knowing the exact protein composition of every cell type under every environmental influence. Since any biological function of the DNA sequence is traditionally expected to be mediated and modified by proteins, it is important to know what mediators are present in the system we investigate. In a simplistic example, a DNA sequence might carry a strong nucleosome positioning signal but if transcription factors that can displace the nucleosome are present, then the theoretical

prediction would be wrong. On the same note, we are still far from knowing the chemical equilibrium values of all the biological interactions. In the previous example this would mean that the transcription factor that can displace nucleosomes should also have a lower dissociation constant for binding to that particular DNA sequence than the nucleosome. If there is no information about these values there is no way of predicting how the system will behave. The study of these phenomena is the subject of "Systems Biology".

The other main reason we can not predict the biological role of DNA sequences, especially regarding transcription, is the influence of the chromatin environment in which these regulatory elements are embedded. In the promoter/enhancer regions for example it is generally agreed that the sequence is the principal regulator that provides the recognition and binding template for transcription factors. But as position effect variegation and transgene silencing phenomena suggest, the influence of these regulatory elements can be overridden if the surrounding chromosomal region dictates so. This happens even in the presence of all the protein components that are necessary for expression. Contemporary research of this all-important "chromatin environment" focuses on histones and their post-translational modifications and DNA methylation. It is the area of study of the field of epigenetics.

1.2. Histone modifications

The eukaryotic genomic DNA is found as a complex with histone proteins. Histones H2A, H2B H3 and H4 are the core histones around which the double strand of DNA is wrapped, while histones H1 and H5 are the linker histones. Two sets of the core histones whith the DNA around them form a nucleosome. The core histones are post-translationally modified at their amino-terminal end that protrudes from the nucleosome. The histone modifications involve methylation, acetylation, phosphorylation, glycosylation, ADP-ribosylation and ubiquitylation and happen at

specific amino acids. The modifications that are the most relevant to the work presented here are going to be outlined here.

Acetylation and de-acetylation of histones H3 and H4 are catalysed by histone acetyltransferases (HATs) and deacetylases (HDACs) respectively. Acetylation and de-acetylation is a very dynamic process, important for the determination of the transcriptional status of the genes they are associated with; acetylation is associated with transcription activation and de-acetylation with repression (Spencer and Davie 1999). In the majority of cases, methylation at the aminoterminal domains of H3 and H4 is catalysed by members of the SET-domain family (Dillon *et al.* 2005). Methylation at H3K4 and H3K36 is indicative of active genes (Bernstein *et al.* 2002; Santos-Rosa *et al.* 2002; Xiao *et al.* 2003; Krogan *et al.* 2003; Ng *et al.* 2003; Bernstein *et al.* 2005), while methylation at H3K9 and H3K20 is associated with gene repression (Aagaard *et al.* 1999; Guanchao Jiang 2004). Another study has shown that H3K9 methylation is also associated with the transcribed region of active genes (Vakoc *et al.* 2005). A special case of histone methylation is H3K27 that is catalysed by polycomb-group proteins and is related with transcriptional repression (Muller *et al.* 2002; Czermin *et al.* 2002; Kunert *et al.* 2003).

1.3. DNA methylation

Histone modifications vary in different chromatin regions and often reflect the methylation status of the DNA in these regions. However, DNA methylation is the only epigenetic modification that directly affects the DNA chain and this raises the possibility that DNA methylation could be an important link between DNA sequence and chromatin function. However, its importance in higher organisms is disputed, as many eukaryotes are virtually devoid of it. The fact that inhibition of DNA methylation in the organisms that normally have it is embryonic lethal, shows that, in the species in which it is present, it plays a vital role for the survival of the individual.

1.3.1. DNA methylation across organisms

DNA in nature can be found methylated at the C-5 or N-4 positions of cytosine and at the N-6 position of adenine (Figure 1-1). In multicellular eukaryotes only C-5 methylcytosine exists. The methyl group of 5'-methylcytosine in the double stranded helix has been shown to protrude to the major groove (Mayer-Jung *et al.* 1997). Methylation of cytosines in the genome is highly dependent on the sequence context. In *Arabidopsis thaliana* and other plant species, methylcytosines can be found in a CpG, CpNpG, or Cp(A/T)p(A/T) context. In mammals, methylcytosine is mainly found in a CpG context, although there is evidence of limited CpA, CpT and CpC methylation at specific developmental stages (Ramsahoye *et al.* 2000; Haines *et al.* 2001; Dodge *et al.* 2002; White *et al.* 2002). Finally, it has been shown that the genome of *Drosophila melanogaster* contains traces of methylcytosine in every dinucleotide (CpN) context (Lyko *et al.* 2000).

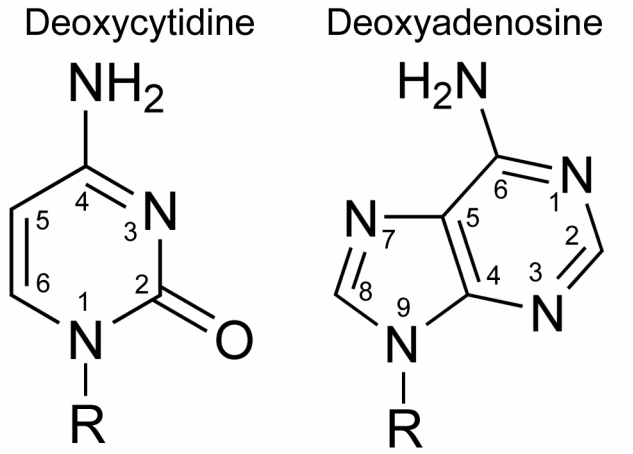

Figure 1-1. Chemical structure of deoxycitidine and deoxyadenosine. R, deoxyribose sugar.

The proportion of the methylcytosines versus all cytosines in a CpG context in the eukaryotic genomes varies considerably from ~0 to 90%. It is interesting that invertebrates seem to be at the lower end of methylation levels, while vertebrates have a high proportion of methylcytosines in their genome. A divide between vertebrates and invertebrates seems to exist also in the way methylcytosines are distributed in the genome. In mammals, methylation is dispersed throughout the genome with the exception of CpG-rich regions, called CpG islands, that are

generally methylation-free and roughly correlate with the promoters of genes. The unusual CpG frequency of CpG islands most likely reflects the AT-rich character of the bulk of the genome due to the hypermutability of methylcytosine to thymine (Bird 1980; Duncan and Miller 1980). On the other hand, the genomes of invertebrates have patchy methylation, with regions that are rich in methylation interspersed in methylation-free DNA (Tweedie *et al.* 1997; Suzuki *et al.* 2007).

There have been several hypotheses that try to explain the differences regarding methylation levels among eukaryotes. One model suggests that methylation is primarily a defence mechanism against repetitive sequences, a function inherited from the bacterial exogenous DNA defence mechanism (Bestor 1990). According to this hypothesis, the differences in methylation levels reflect the genomic content of repetitive sequences. This hypothesis is supported by the observations that some transposable elements are usually found heavily methylated in the genome (Yoder *et al.* 1997) and that lack of methylation causes increased levels of IAP transcription (Walsh *et al.* 1998). Nevertheless, Simmen *et al.* (1999) and Suzuki *et al.* (2007) have shown that the methylated regions of the invertebrate, *Ciona intestinalis* genome, do not show any preference for repetitive sequences. Recent large-scale projects have finally shown that the proportion of the repetitive sequences in the genomes of different organisms does not agree with this being the reason for the differences in methylation levels. In their review on repetitive elements Shapiro and Von Sternberg (2005) show that in *Caenorhabditis elegans* and *Arabidopsis thaliana*, 16.6% and 13-14% of the genome respectively consists of repetitive elements, nevertheless methylation has never been detected in the first. Similarily, in *Mus musculus* and *Drosophila melanogaster* approximately 40% of the genome is non-coding repetitive elements but only traces of methylcytosine have been found in the latter.

An alternative to the genome defence idea is that methylation has evolved to suppress general background transcription (Bird 1995). Although this hypothesis had been initially based on an overestimation of the gene number in vertebrates before the genome projects, recent information has come to light that provides support to this idea. Extensive methylation analysis of *Ciona intestinalis* genomic DNA (Suzuki *et al.* 2007) has showed that methylation coincides with the body of housekeeping

genes. Gene body methylation has also been reported for *Arabidopsis thaliana* (Tran *et al.* 2005; Zhang *et al.* 2006; Zilberman *et al.* 2007) and the insects *Apis melifera* (Wang 2006) and *Myzus persicae* (Field 2000). According to this hypothesis, genes that need to be constitutively expressed at low levels can not afford to allow cryptic promoters in the gene body to interfere with transcription, and methylation is in place to suppress such elements. On the other hand, genes that have strong promoters and are expressed at high levels do not have a need for such a mechanism and remain methylation-free. The model relies on the fitness trade-off between having tight regulation of transcription and a hypermutable methylcytosine-rich genome. Nevertheless, the different levels and patterns of methylation between organisms are not yet explained by this model.

Another theory proposes that the presence of DNA methylation in a multicellular organism depends on its developmental strategy (Jablonka and Lamb 1995). For organisms like *C. elegans* and *D. melanogaster* that have a short life span and little cell turnover it is beneficial to avoid methylation and the high mutation rates that are associated with it. Organisms that go through many cell duplications and an extended life span on the other hand, cannot afford to loose the benefit of a cellular memory that can persist through cell divisions. It is an interesting idea that has been tested for a variety of invertebrates with different developmental strategies and different degrees of methylation (Regev *et al.* 1998). In the study, invertebrate organisms were grouped into categories according to their cell turnover and their methylcytosine content was plotted. This approach showed a very good correlation between low methylation and high cell turnover.

A last model proposes a correlation between body temperature and methylation levels (Jabbari *et al.* 1997). In more detail, this hypothesis sees DNA methylation levels as being inversely proportional with the body temperature of the animal. GC and CpG content follow the same pattern. The authors argue that an increase in temperature causes the methylcytosine to thymidine transition to accelerate, thus selecting for the minimum amount of methylcytosines that are necessary for the proper function of the organism. Further support to this idea comes from studies in fish and reptiles that have different strategies for maintaining their body temperature (Varriale and Bernardi 2006a; 2006b).

1.3.2. DNA methyltransferases

The class of enzymes that is responsible for the methylation of cytosines at the 5' position are cytosine methyltransferases that catalyse the transfer of the methyl-group from S-adenosyl-L-methionine (SAM) to the DNA. Cytosine methyltransferases are present in most eukaryotes (Figure 1-2).

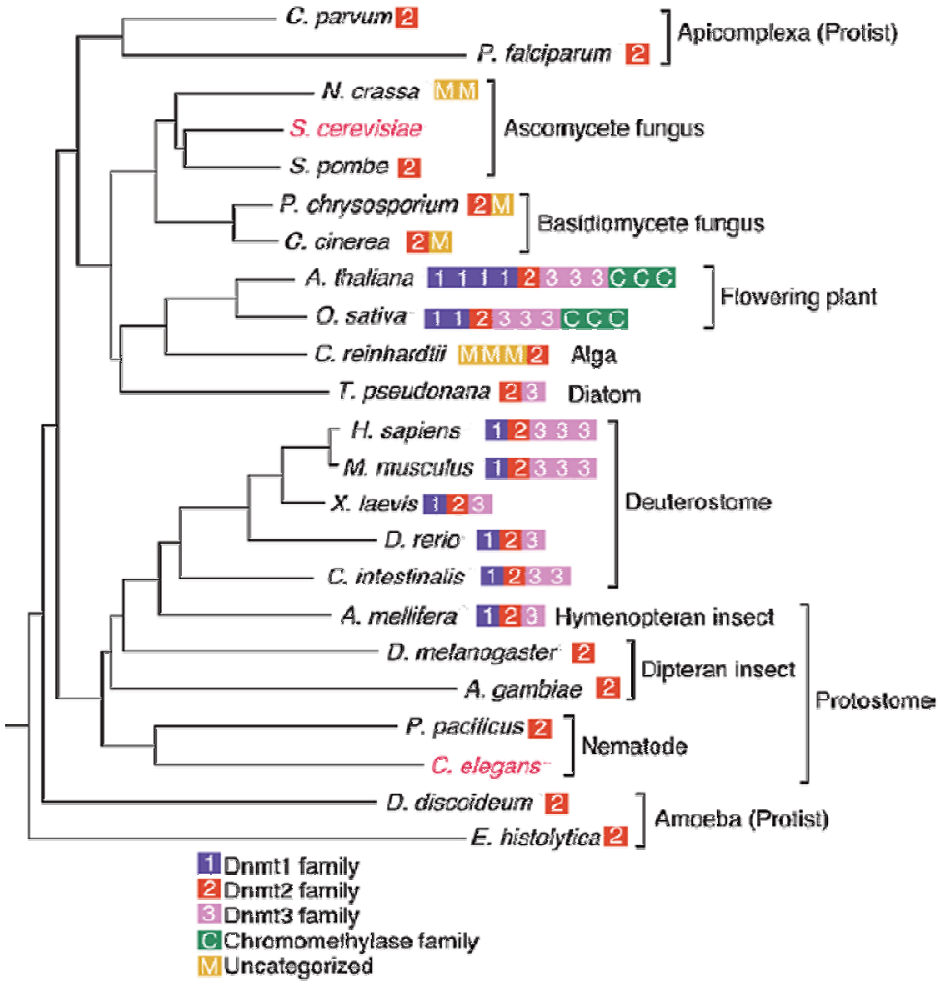

Figure 1-2. Distribution of the various DNA methyltransferases in different eukaryotes. The organisms are classified according to the sequence of 18S rDNA. *S. cerevisae* and *C. elegans* that do not appear to contain DNA methyltransferases are shown in red. (Adapted from Goll and Bestor 2005)

There are four families of DNA methyltransferases, the DNA methyltransferase 1 (DNMT1), DNA methyltransferase 2 (DNMT2), DNA methyltransferase 3 (DNMT3) and chromomethylase families. Chromomethylases

are only found in flowering plants. Mice and humans have one DNMT1, one DNMT2 and three DNMT3 genes (Figure 1-3). They all share some homology at their carboxyterminal catalytic domain although their specificities, and in some cases even their functions, differ. The aminoterminal domain contains sequences that are responsible for protein-protein interactions. Knock-out of the catalytic domain of DNMT1, DNMT3a and DNMT3b in mouse embryonic stem (ES) cells, completely abolished CpG methylation (Tsumura *et al.* 2006). This shows that CpG methylation is exclusively dependent on these enzymes.

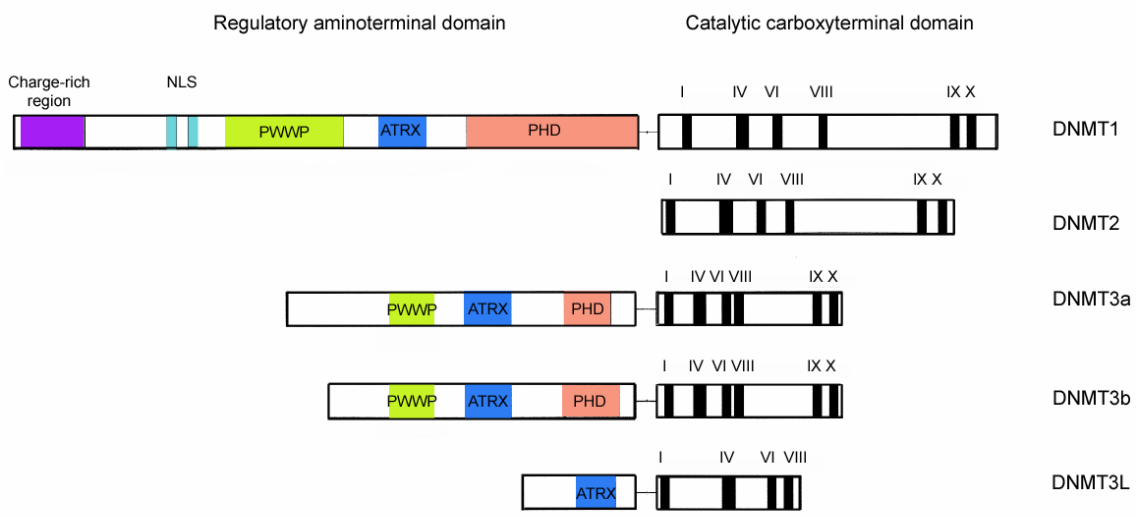

Figure 1-3. Cytosine methyltransferases of mammals. NLS, nuclear localisation signal; PWWP, conserved PWWP tetrapeptide; ATRX, cysteine-rich zinc-finger DNA-binding motif; PHD, polybromo homology domain; I-X, catalytic conserved motifs. (Adapted from Turek-Plewa and Jagodzinski 2005)

DNMT1

DNMT1 catalyses the transfer of methyl-groups in a processive manner (Hermann *et al.* 2004). This means that it transfers many methyl-groups without the release of DNA, something that can be demonstrated experimentally as a lack of intermediate products. DNMT1 is considered to be the maintenance methyltransferase in the cell, as it has been shown that *in vitro* it prefers DNA with hemimethylated 5'-CpG-3':3'-GpC-5' (Fatemi *et al.* 2001). The maintenance activity implies that it is the enzyme responsible for methylating the hemimethylated

CpGs that are present in the double strand of DNA after replication. Indeed, it has been shown to localise at the replication foci during S phase (Leonhardt *et al.* 1992). Its action however is not restricted there, as deletion of the domain that is responsible for the interaction with the replication machinery (charge-rich region, Figure 1-3) showed that DNMT1 activity is independent of this interaction (Spada *et al.* 2007). According to a report by Robertson *et al.* (2000), DNMT1 associates with the E2F-Rb complex that acts to silence genes *de novo*. DNMT1 also interacts with the histone deacetylases 1 and 2 (HDAC1 and 2), the H3K9 methyltransferase SUV39H1, heterochromatin protein 1 (HP1) and the chromatin remodeler SNF2 (Robertson *et al.* 2000; Fuks *et al.* 2000; Rountree *et al.* 2000; Fuks *et al.* 2003a). In summary, the experiments suggest that DNMT1 has a role in both maintenance of the methylation patterns and *de novo* epigenetic modifications in the genome.

There are three DNMT1 isoforms, DNMT1s, DNMT1o and DNMT1p that are found in somatic tissues, oocytes and pachytene-stage spermatocytes respectively (Mertineit *et al.* 1998). They are all products of alternative splicing that causes either an increase of the size of the first exon of the DNMT1p transcript or a small decrease of that of DNMT1o. DNMT1o is present throughout oogenesis and during early embryogenesis until the blastocyst stage and is actively retained in the cytoplasm despite its strong nuclear localisation signal. It enters the nucleus at a specific developmental stage that appears to be determined by the embryo's molecular clock (Cardoso and Leonhardt 1999). Post-translation modifications that have been observed to alter the electrophoretic mobility of DNMT1o in comparison to DNMT1s could be playing a role in the timely entry to the nucleus (Carlson *et al.* 1992). The transition from the DNMT1s to the DNMT1p transcript happens at the early pachytene stage of spermatogenesis and results in repression of DNMT1p translation (Mertineit *et al.* 1998). It seems that a common mechanism involving elimination of DNMT1 from the nucleus via alternative splicing is employed during gametogenesis and early developmental stages. The implications of this will be discussed later.

DNMT2

Until recently, DNMT2 posed a mystery because it is in many cases the only cytosine methyltransferase present in organisms that are virtually free of methylation,

such as *D. melanogaster* and *S. pombe* (Figure 1-2). Furthermore, the cytosine methyltransferase activity of DNMT2 *in vitro* is very low and it localises to the cytoplasm. DNMT2 overexpression in *Drosophila*, caused a significant increase in cytosine methylation in a Cp(A/T) context (Kunert *et al.* 2003). This indicated that DNMT2 might have a role for DNA methylation establishment in *Drosophila*. However, only trace amounts of Cp(A/T) methylation exist in this organism. A satisfactory explanation for the role of DNMT2 has been recently provided by Goll *et al.* (2006). These researchers showed that DNMT2 in mouse, *Drosophila* and *A. thaliana* is actually an aspartic acid transfer-RNA ($tRNA^{Asp}$) methyltransferase. Apparently, in this case, evolution employed the methyl-transferring activity of the enzyme in another biological function.

DNMT3

Mammals have three DNMT3 enzymes, DNMT3a and DNMT3b have *de novo* methyltransferase activity, while DNMT3L lacks methyltransferase activity and is a co-factor of DNMT3a and b, expressed during gametogenesis (Hata *et al.* 2002). DNMT3L probably acts by facilitating SAM loading onto the DNMT3 enzymes (Kareta *et al.* 2006) and, as it will be discussed later, it has a role in the establishment of parental imprints. The *de novo* methyltransferase activity of the other two enzymes means that they can methylate unmethylated substrates. *In vitro*, DNMT3a has been shown to prefer the DNA strand in which the CpG site is flanked by pyrimidines leading to hemimethylated products (Lin *et al.* 2002). This, together with the observation that the DNMT3 proteins co-immunoprecipitate with DNMT1 (Kim *et al.* 2002; Rhee *et al.* 2002), gives rise to the possibility that DNMT3s act by hemimethylating their substrate and DNMT1 is responsible for completing the reaction. However, verification of this begs further experimental evidence. The biochemical characteristics of the two enzymes differ; DNMT3b is processive (*i.e.* methylates more than one cytosines without releasing the DNA), while DNMT3a is distributive (*i.e.* methylates one cytosine at a time) (Gowher and Jeltsch 2002). The activity of DNMT3a on nucleosomal DNA is much lower that that of DNMT3b (Takeshima *et al.* 2006) and the opposite is true for naked DNA (Suetake *et al.* 2003). This could mean that DNMT3a and DNMT3b *in vivo* act on naked and nucleosomal DNA respectively.

Despite their distinct biochemical properties, the particular roles of the two enzymes *in vivo* are not very well understood. Deletion of the catalytic activities of one or the other enzyme showed that, at the majority of the studied loci, methylation was not affected (Okano *et al.* 1999). When both the enzymes were knocked out, and only then, could the methylation be erased in these loci. This means that, in most cases, the two enzymes probably complement each other. This is further supported by the fact that DNMT3a and b co-immunoprecipitate both *in vivo* and *in vitro* (Kim *et al.* 2002). In the study of Okano *et al.* (1999) DNMT3b alone was shown to be responsible for the methylation of centromeric minor satellite repeats. DNMT3a and not DNMT3b on the other hand appears to be able to restore the methylation of the *Xist* and *H19* genes in DNMT3a/b -/- cells (Chen *et al.* 2003). It is still unknown if the two enzymes have different specificities in the case of other genes.

Like DNMT1, DNMT3a and DNMT3b interact with other proteins. In more detail, DNMT3L has been shown to associate with histone H3 tails that are not methylated at lysine 4 and to induce methylation of the associated DNA (Ooi *et al.* 2007). In addition, DNMT3L forms a symmetric heterotetramer with DNMT3a, altering its structure and puting constraints on the distance of CpGs that can be methylated by the complex (Jia *et al.* 2007). DNMT3a also reportedly associates with HDAC1 (Fuks *et al.* 2001; Bai *et al.* 2005), the H3K9 methyltransferases G9a (Feldman *et al.* 2006), SUV39H1 (Fuks *et al.* 2003a) and SETDB1 (Li *et al.* 2006a) and with HP1 (Fuks *et al.* 2003a), while DNMT3b has been shown to interact with HDAC2 (Bai *et al.* 2005). Additionally, both DNMT3a and DNMT3b have been shown to associate with the EZH2 polycomb-group protein at repressed genes (Vire *et al.* 2006) while another polycomb-group protein, Cbx4, might regulate DNMT3a activity through SUMOylation (Li *et al.* 2007). All these interactions provide a mechanistic explanation for the observed relationship between heterochromatin and methylated DNA. Functionally, the most important interaction seems to be that with HDACs since HDAC association alone has been shown to be enough for trichostatin A (TSA)-sensitive repression of a reporter gene by DNMT3a (Fuks *et al.* 2001). Moreover, DNMT3b-associated HDAC activity regulates differentiation of pheochromocytoma cells (Bai *et al.* 2005).

DNMT3a has three isoforms that are all enzymatically active. Isoforms DNMT3a-α and-β (Weisenberger *et al.* 2002) are products of alternative splicing at exon 1 and they are often considered together as DNMT3a. DNMT3a-α is the most abundantly expressed of the two in all tissues except testes, where DNMT3a-β seems to take over. The third isoform, DNMT3a2 is the product of an alternative promoter that is active in ES but not fibroblast cells (3T3) (Chen *et al.* 2002). DNMT3a2 misses regions of the aminoterminal domain and this appears to cause a more diffuse localisation at the nucleus. DNMT3a2 is ubiquitously present in all examined tissues with the exception of brain, in which it can not be detected.

There are six isoforms of DNMT3b (DNMT3b1 to b6), all produced by alternative splicing (Robertson *et al.* 1999). Of all the isoforms, only DNMT3b1 and 2 appear to be catalytically active (Aoki *et al.* 2001; Chen *et al.* 2003). The other four isoforms are missing parts of the catalytic carboxyterminal domain. They are nevertheless all expressed in different combinations in the tissues tested (Robertson *et al.* 1999). Regarding the DNA binding capacity of the inactive isoforms, the subnuclear localisation to heterochromatin is lost in DNMT3b3 (Ueda *et al.* 2006). Additionally, DNMT3b3 and b6 can be depleted after 5'-azacytidine treatment which indicates they can interact with DNA (Weisenberger *et al.* 2002). This method relies on the covalent trapping of the enzyme to the DNA where the nucleotide analogue has been incorporated. It is a not immediately clear why the cell would invest resources on transcribing the inactive forms of DNMT3b. A possibility is that they could be active through their aminoterminal domain or perhaps their transcription is a means to regulate DNMT3b levels without altering the transcription rate.

Finally, DNMT3L was recently found to exist in three isoforms produced from alternative promoters (Shovlin *et al.* 2007). The typical promoter seems to be active in stem cells and early spermatogenesis, while at late pachytene spermatocytes a second promoter is activated that causes the transcription of a truncated, non-coding mRNA. Finally, a third promoter is active in the oocytes and causes transcription of a longer, active form of DNMT3L. The implications that these discoveries might have during differentiation is discussed later.

1.3.3. Establishment of DNA methylation patterns

As already discussed (section 1.3.1), the DNA methylation profile is very different between vertebrates and invertebrates genomes. In the first, the entire genome is globally methylated with the exception of CpG islands. In the second, there are methylated regions that are interrupted by long unmethylated regions. Furthermore, methylation in vertebrates can be found at inactive promoters but transcriptional activity cannot predict the methylation status of the 5' upstream region of a gene. What determines these methylation patterns? What's more, is there an active targeting mechanism that selects some regions and not others for methylation?

In vitro experiments have not revealed any intrinsic sequence specificity of the DNMT enzymes. Specificity however, could be conferred indirectly through interactions with transcription factors. Transcription factors are known to display sequence specificity and DNMTs have been shown to associate with E2F-Rb (Robertson *et al.* 2000), GCNF (Sato *et al.* 2006), COUP-TF1(Gallais *et al.* 2007), PML-RAR (Di Croce *et al.* 2002) and RP58 (Fuks *et al.* 2001). Most probably the list is still incomplete and this is potentially an effective way for the determination of the methylation patterns of the regulatory elements of genes.

Another mechanism for determining the methylation pattern through a transcription factor has been reported for the *aprt* gene (Brandeis *et al.* 1994; Macleod *et al.* 1994; Mummaneni *et al.* 1998). In the promoter of the active gene, Sp1 binding to its response element protects the region from methylation. Evidence for a similar mechanism for the prevention of DNA methylation has been recently uncovered for the imprinted *H19/Igfr2r* locus (Engel *et al.* 2006). In this study, binding of the CTCF factor to the differentially methylated region (DMR) of the maternal allele seems to prevent methylation and regulate enhancer activity in *cis*.

The examples examined above concern establishment of DNA methylation patterns at the regulatory elements of genes but there is no easy way to explain the genome-wide methylation observed in vertebrates through the mediation of transcription factors. A mechanism that seems to satisfactorily explain the methylation at the repetitive elements of flowering plants is RNA-directed DNA methylation (RdDM) (Wassenegger *et al.* 1994; Mette *et al.* 2000). In this case,

double-stranded antisense RNA transcripts are recognised by the RNAi core enzyme Dicer and cleaved to 21-24mers. Through a less well understood mechanism, these small interfering RNAs (siRNAs) cause methylation of the DNA that is homologous to them. It is an elegant mechanism that takes advantage of the RNA:DNA complementarity. An RNAi-related mechanism has also been implicated in the heterochromatinisation of centromeric and other repeats in many organisms that do not have DNA methylation.

The existence of an RdDM mechanism in mammals that is analogous to that of plants is controversial. On one hand, RNA polymerase VI (Wassenegger and Krczal 2006) and chromomethylases (Cao *et al.* 2003) are central to RdDM in plants but neither of these enzymes has homologues in animals. Similarily, CpNpG methylation is typical of RdDM but only traces of non-CpG methylation can be detected in animals (Ramsahoye *et al.* 2000; Haines *et al.* 2001; Dodge *et al.* 2002; White *et al.* 2002). In an older study however, CpNpG methylation occurred quite frequently in a stably transfected plasmid (Clark *et al.* 1995). In this case methylation levels reached 20-40% for the CpA/TpG trinucleotide. It would be interesting to confirm this result.

Importantly, in human, Tufarelli *et al.* (2003) showed that a chromosomal deletion in a patient of α-thalassemia caused transcription of an antisense RNA of the *α2-globin* gene. This antisense RNA was responsible for dense methylation and repression of the *α2-globin* gene. This phenomenon is reminiscent of the *Xist* and *Air* genes, in which regulated transcription from an antisense promoter causes allele-specific methylation. In all the cases however, antisense transcripts are several Kb long and the RNAi machinery does not seem to be involved. The best indication that RdDM might be acting in mammals are the experiments by Kanellopoulou *et al.* (2005), in which conditional knock-out of the Dicer gene resulted in reduced methylation in centromeric and pericentromeric repeats. Unfortunately, until more specific examples have been discovered, evidence for RdDM in mammals remains elusive.

A methylation pattern that seems to be shared between animals and plants is gene body methylation (Field 2000; Tran *et al.* 2005; Zhang *et al.* 2006; Zilberman et al. 2007; Suzuki *et al.* 2007). In more detail, in invertebrates DNA methylation

seems to be confined in the gene body of housekeeping genes (Suzuki *et al.* 2007), while in plants high gene body methylation levels appear to correlate with transcription (Zilberman *et al.* 2007).

There is no experimental evidence to my knowledge that sheds light on the mechanism that is responsible for this methylation pattern. It is difficult however to ignore the similarity with methylation at lysine 36 of histone 3 (metH3K36), that has also been shown to localise at the gene-body of actively transcribed genes. In these cases, at least in yeast, metH3K36 is brought about because its specific methyltransferase, Set2, associates with the elongating form of RNA polymerase II (Krogan *et al.* 2003; Xiao *et al.* 2003). Set2 also recruits histone deacetylases that are deacetylating the gene body (Krogan *et al.* 2003; Joshi and Struhl 2005). Given the strong evidence for the association of DNMTs with histone deacetylases (section 1.3.2), it is not unreasonable to imagine that DNMTs are recruited to the gene body through their association with HDACs. Unfortunately the genome of *S. cerevisiae* is not methylated. It would be interesting to show if the same H3K36 methylation mechanism exists in other organisms that have methylation. If this is the case, then gene-body methylation should be happening downstream of H3K36 methylation.

1.4. Proteins that read the methylation signal

The identification of proteins that can read the methylation signal is imperative for our understanding of both how DNA methylation has evolved in eukaryotes and what its function is. Our current knowledge about this first step of how the DNA methylation pattern communicates information to the cell is outlined here. A comparison of the domains of the proteins described here with regard to their function is shown in Figure 1-4.

1.4.1. The methyl CpG-binding domain (MBD) family

The best characterised methylcytosine-binding protein family is the MBD family. It consists of five members, namely MeCP2, MBD1, MBD2, MBD3 and MBD4, which are expressed in most of the tissues tested (Roloff *et al.* 2003). They

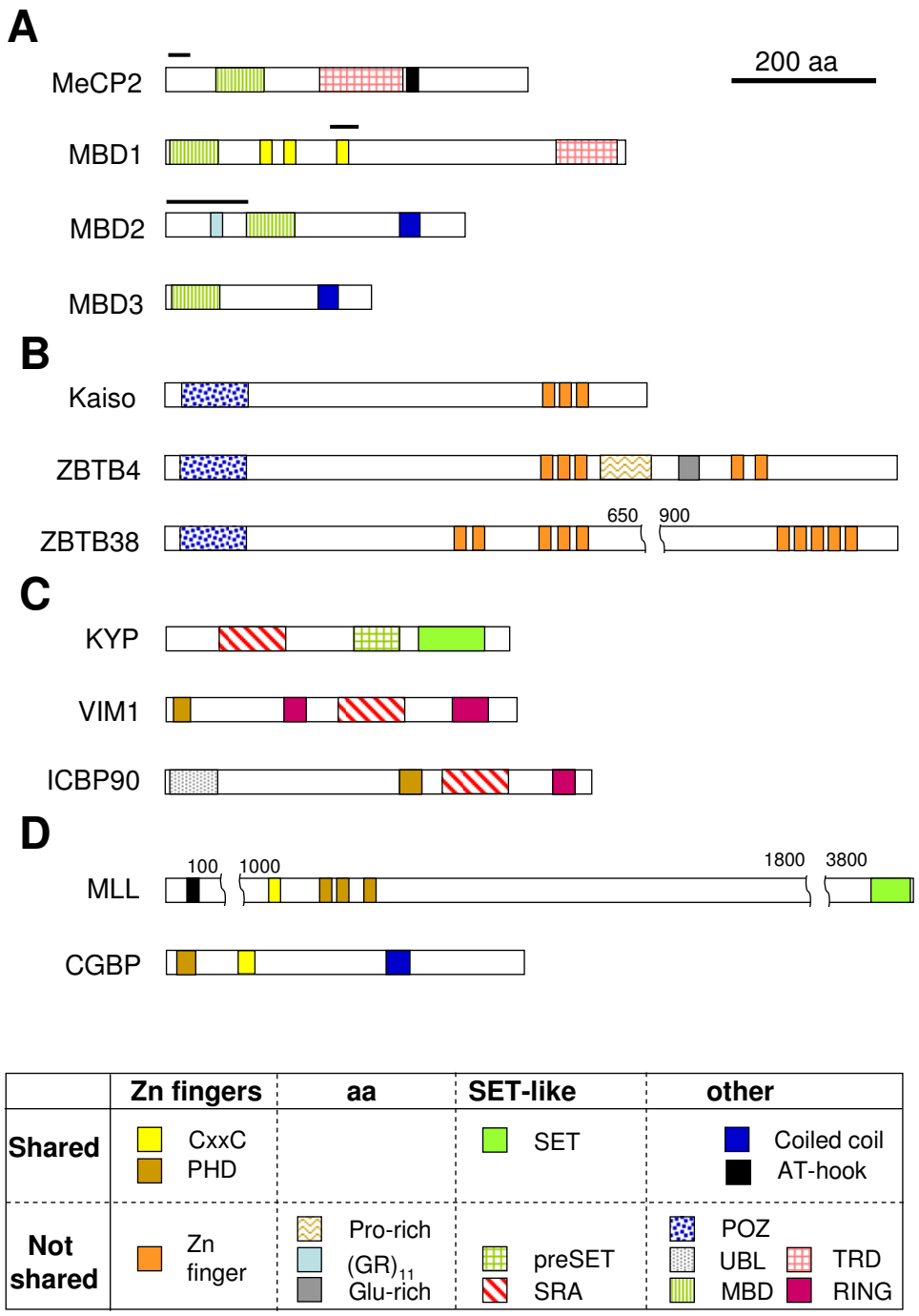

Figure 1-4. Proteins that read the methylation signal. Graphical representation of the domains in the proteins that recognise and bind specifically methylcytosines (A,B and C) and cytosines (D). MBD 1 is a special case in that it can recognise both. (A) MBD family. (B) Kaiso and related proteins. (C) SRA-domain proteins. (D) MLL and CGBP. Black lines above the proteins show the regions excluded in alternative transcripts. The domains, organised according whether they are shared or not by more that one categories of the proteins shown here, are shown at the bottom.

are paralogues that all have a conserved MBD domain and share a maximum 63.8-94 % similarity (Roloff *et al.* 2003). The MBD domains of these proteins, except MBD3, bind specifically methyl-CpGs. NMR analyses of the MBD domain of MeCP2 have shown that it forms a wedge-like structure of four antiparallel β-sheets which fits into the major groove of the DNA with a positively charged surface andbinds the protruding methyl-group (Wakefield *et al.* 1999; Ohki *et al.* 1999).

In general, MBDs are transcriptional repressors with little or no overlap of their target genes (Klose *et al.* 2005). As it will be discussed in more detail later, they are known to associate with various histone modification enzymes and chromatin remodelling complexes. This shows the close link between transcriptional status, DNA methylation and chromatin. The causal relationships however remain to be shown.

The only MBD protein that is not a transcriptional repressor is MBD4, which is a mismatch repair enzyme. MBD4 is a DNA glycosylase that preferentially binds to methylCpG:TpG mismatches produced by spontaneous methylcytosine deamination (Hendrich *et al.* 1999; Wu *et al.* 2003; Millar *et al.* 2002; Petronzelli *et al.* 2000). A recent report has shown that MBD4 has HDAC- and DNA methylation-dependent repressor activity (Kondo *et al.* 2005). Because of the limited evidence for a methylation dependent role of MBD4 however, it will not be considered here.

MeCP2

MeCP2 is an X-linked gene that has two isoforms, *MeCP2α* and *MeCP2β* that are the products of alternative splicing of exon 2 (Kriaucionis and Bird 2004). Of the two isoforms, the shorter *MeCP2α* is more abundant in tissues and is translated more efficiently. MeCP2 null mice have neurological defects and limited viability (Guy *et al.* 2001) as well as mitochondrial abnormalities (Kriaucionis *et al.* 2006). In humans, mutations in MeCP2 cause the autism spectrum disorder disorder Rett syndrome (Van den Veyver and Zoghbi 2001). The aminoterminal region of MeCP2, which also contains the MBD domain, co-immunoprecipitates with an H3K9 methyltransferase activity (Fuks *et al.* 2003b). Other reported interactions of MeCP2 include, but are not confined to, DNMT1 (Kimura and Shiota 2003) and HDACs (Nan *et al.* 1998). The latter seem to be in the context of the Sin3a chromatin remodelling complex. Nevertheless, detailed analysis has shown that MeCP2 is

usually found as a monomer in the cell and probably the above interactions are only transient (Klose and Bird 2004).The observation that MeCP2 localises to heterochromatic foci (Brero *et al.* 2005), indicates a general role in heterochromatinisation. Nevertheless, specific targets of MeCP2 include *H19* (Fuks *et al.* 2003b), *xHairy2a* (Stancheva *et al.* 2003), *Bdnf* (Martinowich *et al.* 2003) and *Dlx5/6* (Horike *et al.* 2005; Klose *et al.* 2005).

MeCP2 *in vitro* and *in vivo* binds to a single methyl-CpG with an AT-run in close proximity through its MBD domain (Klose *et al.* 2005). The transcription repression domain (TRD) which is necessary for transcription repression is also responsible for some of the protein-protein interactions described above. MeCP2-induced transcription repression is sensitive to TSA and it is probably caused through histone deacetylation (Nan *et al.* 1998; Jones *et al.* 1998).

MBD1

Absence of MBD1 in mice causes neuronal defects (Zhao *et al.* 2003). MBD1 is unique among the other MBDs in that it contains CxxC domains. CxxC domains are zinc-finger DNA binding domains present in a variety of DNA-binding proteins. There are four MBD1 (a to d) isoforms, all products of alternative splicing (Fujita *et al.* 1999). MBD1b and MBD1d are missing the third CxxC domain. This, but not the other two CxxC domains, has been shown to target MBD1 to non-methylated DNA both *in vivo* and *in vitro* independently of the methyl-binding specificity of the MBD domain (Jørgensen *et al.* 2004). The dual specificity of MBD1 is intriguing. A recent study has shown that in transient transfection assays MBD1 associates with the PLM-RARα transcription factor, which is an oncoprotein, and is recruited to the promoter of the target *RARβ2* gene (Villa *et al.* 2006). This is independent of methylation and is followed by HDAC recruitment to the promoter and establishment of DNA methylation. At the end of this process MBD1 occupancy seems to have spread over the gene. One can imagine that in this example the CxxC domain facilitates the initial binding to the unmethylated promoter and the MBD domain takes over after methylation has been established. This is nevertheless only a hypothesis and more experimental evidence is required to establish the roles of the CxxC and MBD domains of MBD1.

MBD1 is reported to interact with the H3K9 methyltransferase SUV39H1, HP1, as well as with HDAC3 (Fujita *et al.* 2003; Villa *et al.* 2006) . MBD1 associates also with another H3K9 methyltransferase, SETDB1 (Sarraf and Stancheva 2004). SETDB1 association is important for the MBD1 repressor activity and is interrupted by sumoylation of MBD1 (Lyst *et al.* 2006). Finally, MBD1:SETDB1 forms a complex with CAF-1 during replication, probably contributing to the heritable maintenance of the chromatin's epigenetic information after DNA replication (Sarraf and Stancheva 2004).

MBD2 and MBD3

MBD2 and MBD3 are very closely related proteins (Roloff *et al.* 2003) that have divided from their common phylogenetic ancestor recently, probably after the separation of invertebrates and chordates. Indeed, the invertebrates *C. intestinalis* and *D. melanogaster* seem to have one MBD2/3 protein. It is assumed that MBD2/3 in *D. melanogaster* binds specifically to methylCp(T/A) (Marhold *et al.* 2004). MBD2 is characterised by an arginine-glycine (RG) rich domain that, as it will be explained later, is important for the regulation of the protein. An alternative splicing isoform of MBD2, MBD2b, does not contain this region (Hendrich and Bird 1998).

Despite their close phylogenetic relationship, MBD2 and MBD3 in mammals are very different proteins. Although they both have a conserved MBD domain, MBD3 binding to DNA is independent of methylation, whereas MBD2 binds specifically to methylcytosines (Hendrich and Bird 1998). Moreover, MBD3 -/- mice die during early development, while MBD2 -/- mice are viable and healthy and show only small behavioural abnormalites (Hendrich *et al.* 2001). Closer study of MBD2 -/- mice has revealed that they exhibit abnormal expression of pancreatic enzymes in the colon (Berger *et al.* 2007) and also have a phenotype of increased tumour resistance in the $Apc^{Min/+}$ background (Sansom *et al.* 2003).

Regarding their interaction with other proteins, both MBD2 and MBD3 have been shown to associate with HDAC1 and 2 (Saito and Ishikawa 2002; Le Guezennec *et al.* 2006). Importantly both methylcytosine binding and HDAC association of MBD2 is impaired when the RG-rich aminoterminal region is methylated by PRMT5 (Tan and Nakielny 2006). This seems to be a regulatory mechanism for MBD2 action. In an analogous manner, the protein MBD3L (which is

not part of the MBD family) is speculated to have a role on the regulation of MBD3 (Jin *et al.* 2005). Both MBD2 and MBD3 are core components of the Mi-2/NuRD complex (Ng *et al.* 1999; Zhang *et al.* 1999; Wade *et al.* 1999; Feng and Zhang 2001). Nevertheless, a recent study has demonstrated that MBD2 and 3 do not coexist in the Mi-2/NuRD complex and the presence of one in the complex excludes the other (Le Guezennec *et al.* 2006).

1.4.2. Kaiso

Kaiso binds to sequences with multiple methyl-CpGs and this binding depends on its zinc-finger domain (Prokhortchouk *et al.* 2001). Similar to MBD2 -/- mice, *Kaiso* -/- mice are viable and fertile and have a phenotype of resistance to intestinal cancer in the $Apc^{Min/+}$ background (Prokhortchouk *et al.* 2006). Surprisingly, in *Xenopus laevis*, *Kaiso* depletion caused severe developmental abnormalities and apoptosis (Ruzov *et al.* 2004). There is no adequate explanation for this difference between organisms. *Kaiso* is particularly interesting because it is the only methylation-dependent repressor that is known to be part of signalling cascades; it associates with p120 catenin (Daniel and Reynolds 1999) and seems to be important for Wnt signalling (Kim *et al.* 2004; Park *et al.* 2005). It is also part of the N-CoR (Yoon *et al.* 2003) chromatin remodelling complex. Further study of *Kaiso*-mediated repression could provide the first clues about how environmental signals and epigenetic phenomena cooperate to regulate gene transcription.

Two more proteins have been discovered recently in humans that share significant similarities with Kaiso (Filion *et al.* 2006). They are the ZBTB4 and ZBTB38 proteins which contain both a POZ domain and Kaiso-like conserved zinc-fingers. They are expressed in adult tissues but not embryos. They bind to methylated DNA both *in vivo* and *in vitro* and reporter assays have shown that they are methylation-dependent transcription repressors.

1.4.3. SET- and RING-associated (SRA) domain-containing proteins

KRYPTONITE (KYP) is a protein of *A. thaliana* known for its H3K9 methyltransferase activity. It was recently shown that it can also specifically bind to methylcytosines in every dinucleotide context (Johnson *et al.* 2007). *In vivo*, it

associates with the methylated *AtSN1* and *AtCOPIA* retrotrasposons. The same study showed that the SRA domain of the protein was mainly responsible for its methyl-binding activity.

Investigation demonstrated that another plant histone methyltransferase, SUV6, and two previously uncharacterised proteins, ORTH1 and 2, that also contain SRA domains could specifically bind methylcytosines *in vitro* too (Johnson *et al.* 2007). Of these, SUV6 did not show any binding specificity towards the sequence context of the methylcytosine, while ORTH1 and 2 seemed to prefer CpGs. There is no evidence yet for *in vivo* methylcytosine-specific binding of these proteins.

Another SRA domain protein in *A. thaliana* recently identified by its methyl-binding activity is VARIANT IN METHYLATION 1 (VIM1) (Woo *et al.* 2007). This protein binds specifically to methylcytosines in a CpG or CpNpG context *in vitro* and is associated with centromeric repeats, as well as the tandemly repeated 5S rRNA genes, the repetitive *Athila* retrotransposable elements and a *Cinful*-like retrotransposon *in vivo*.

The only mammalian protein that shows some similarity to the SRA-domain proteins of *A. thaliana* and is reported to have methylcytosine specificity is the human ICBP90. Another protein in mouse, Np95, also has a SRA domain and has a role in centromere heterochromatinisation (Papait *et al.* 2007), but has never been shown to bind methylated DNA. ICBP90 was first identified as a CCAAT binding protein contributing to the regulation of *topoisomerase IIα* and showed distinct expression patterns in different tissues (Hopfner *et al.* 2000). Further investigation, however, has shown some affinity of ICBP90 for methylated CpGs and that it could have a role in methylation-dependent transcription regulation of certain tumor suppressor genes (Unoki and Nakamura 2003; Unoki *et al.* 2003). Nonetheless, its exact role as a methylation-dependent transcriptional factor still remains to be shown.

In summary, the SRA domain proteins are a new class of methylcytosine-binding proteins. In plants, they contain a SET domain and previous studies had shown a role for the heterochromatinisation of DNA by catalysing histone modifications, but it is only now that their DNA-binding ability has been shown. Additionally, they are the only methylcytosine-binding proteins that seem to have a

specific role on methylated repetitive elements found at centromeres. These observations dictate that they could be mediating the effect of DNA methylation on chromosome structure. Nevertheless, more experiments are needed to establish this function.

1.4.4. A link between DNA methylation and histone modifications

Heterochromatin is typically defined as chromatin rich in DNA methylation, deacetylated histones and repressive histone marks such as H3K9 methylation. As described previously (section 1.3.2), DNMTs associate with the enzymes that are responsible for the histone modifications that are present in heterochromatin providing a spatial link between DNA methylation and histone modifications. Additionally, the evidence outlined above show that the enzymes that are responsible for repression of the methylated promoters are also associating with histone modifying enzymes. In the case of MBDs and Kaiso this happens through their interaction with HDACs and H3K9 methyltransferases. The SRA domain-containing proteins on the other hand contain themselves a SET or SET and RING associated motif. The SET motif is typical of histone methyltransferases and indeed, at least in the case of KYP, it has been shown to be able to catalyse H3K9 methylation. It seems that methyl-binding is functionally connected with H3K9 methylation and in the case of SRA proteins evolution has merged the two functions into one enzyme.

I would here like to mention an experiment that emphasises the link between H3K9 methylation and DNA methylation. Lyko *et al.* (1999) showed that expression of mouse DNMT enzymes in *Drosophila* forced establishment of DNA methylation and had dramatic effects on its phenotype. This happened despite the fact that this organism, naturally lacking DNA methylation, does not have the machinery that can read the methylation signal in order to suppress a gene or alter the histone modifications. They later showed that togther with DNA methylation, overexpression of DNMTs caused increase in the global levels of H3K9 methylation.

As described already and will be discussed more extensively later, DNA methylation has a role in transcription repression. A simple model would suggest that DNA methylation attracts the methylcytosine binding transcription repressors. Nevertheless, repression by these proteins is dependent on the histone modification

machinery that is recruited by both the DNA methyltransferases and the repressors themselves. All these interactions (Figure 1-5) form a feed-forward loop of DNA methylation and histone modifications. The causal relationships of these interactions as well as their relative importance is the subject of intense research. The fact that organisms such as *D. melanogaster, C. elegans* and *S. serevisae*, which do not have DNA methylation machinery, have nevertheless a complete set of histone modifying enzymes, suggests that histone modifications likely higher in the hierarchy of epigenetic modifications for the determination of the transcriptional fate of the gene.

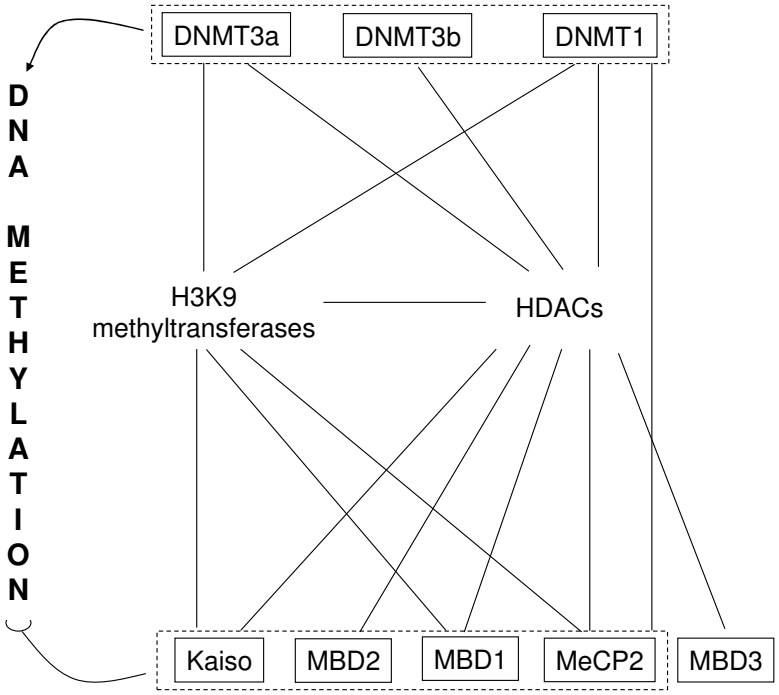

Figure 1-5. Schematic diagram of the known interactions of methylcytosine binding proteins. The diagram focuses on the interactions with the histone modification machinery. Interactions are shown with a line. The methylcytosine binding proteins have a repressive effect on methylated DNA. DNMTs are methylating DNA. The proteins that have the same effect on DNA are grouped together.

1.4.5. Proteins that bind unmethylated CpGs

The methylation pattern in the genome is made up of both methylated and unmethylated DNA. The previous sections focused on the proteins that can read

methylation. This section describes the two known proteins in mammals that specifically bind non-methylated CpGs.

CpG-binding protein (CGBP)

CGBP (also known as CXXC finger protein 1 (CFP1)) is a transcriptional activator (Shin Voo *et al.* 2000) that specifically binds unmethylated CpGs through a CxxC domain that recognises a single CpG in a (A/T)pCpGp(A/T) context (Lee *et al.* 2001). CGBP -/- mice die around implantation indicating an important role of the protein for development (Carlone and Skalnik 2001). Interestingly, CGBP -/- ES cells show reduced methylation levels in both repetitive elements and gene-associated sequences, which is considered to be a secondary effect of the observed DNMT1 downregulation (Carlone *et al.* 2005). Recent experiments have shown that CGBP is also involved in the methylation of histone H3 at lysine4 (K4) (Lee and Skalnik 2005).

Mixed lineage leukaemia (MLL)

Another protein that is specific for unmethylated CpGs is MLL. The region that is responsible for this specificity appears to be a CxxC-like domain (Birke *et al.* 2002). Site selection experiments have shown that this protein prefers CpG-rich stretches of DNA rather than single CpGs (Birke *et al.* 2002), making it a good candidate for a CpG island-specific transcription factor. Indeed, all of the genes that MLL has been shown to associate with, have CpG islands (Milne *et al.* 2005). An interesting class of genes that depends specifically on MLL for their expression is the *Hox* family (Terranova *et al.* 2006) and the misregulation of these genes explains satisfactorily the homeotic phenotype of the MLL mutant mice. Finally, MLL is also an H3K4 methylatransferase and the SET domain at its carboxyterminal end is responsible for this activity (Milne *et al.* 2005; Nakamura *et al.* 2002).

1.5. Roles of DNA methylation

1.5.1. Role in transcription

As discussed extensively above (section 1.4), DNA methylation at the promoter has a negative effect on transcription because of the repressive action of methylcytosine binding proteins, such as the MBDs and Kaiso. The situation is nevertheless not so straightforward, as experiments suggest that, at least in the cases studied, DNA methylation may be the result rather than the cause of transcriptional repression. Early studies on the time course of transcriptional repression of an *in vitro* methylated gene injected into *X. laevis* oocytes (Kass *et al.* 1997), have shown that transcription is not affected by the presence of methylation in the promoter but by the assembly of repressive nucleoprotein complexes. More recently, Mutskov and Felsenfeld (2004) have shown that a transgene's expression levels were closely mirrored by a reduction in histone acetylation and H3K4 methylation but DNA methylation at the promoter was established more gradually as repression proceeded. A similar situation has been observed in the endogenous *Oct4* gene in early development (Feldman *et al.* 2006; Gu *et al.* 2006) and will be discussed in more detail in the third chapter.

Another mechanism by which promoter methylation affects transcription is by forbidding the binding of transcription factors that are necessary for the activation of the gene. Examples of such a mechanism include Sp1 binding at the *Abcc6* promoter (Douet *et al.* 2007) and E2F binding at the promoters of *dihydrofolate reductase*, *E2F1*, *cdc2*, *c-myc* and *c-myb* (Campanero *et al.* 2000). An observation that could be related to methylation-sensitive transcription factor binding is that the presence of methylcytosines might cause conformational changes on the DNA, potentially preventing binding (Muiznieks and Doerfler 1994).

Additionally, it seems that DNA methylation might have an effect on RNA polymerase II itself. Lorincz *et al.* (2004) have demonstrated that methylation of a transgene resulted in a specific decrease of the elongating form of RNA polymerase II occupancy in the methylated region. The authors argue that this might be an indirect effect of the accompanying repressive histone modifications. Finally,

misregulation of transcription due to abnormal methylation patterns is known to happen in cancers, in which, part of the aberrant transcription phenotype can be attributed to methylation.

1.5.2. Role in imprinting

A specialised role of DNA methylation in transcriptional repression is the case of imprinted genes. Imprinting is the phenomenon in which a gene is monoallelically expressed depending on whether it is of maternal or paternal origin. In a simplified model, parent-specific methylation of the imprinted gene is determining its transcription status. Although the purpose of such a mechanism in higher organisms is not known, the most popular current theory proposes that it is the product of competition for resources during embryogenesis. In more detail, this "conflict hypothesis" (Moore and Haig 1991) speculates that imprinting is the product of evolution forces acting on genes that affect embryonic development. In more detail, in organisms that rely on maternal resources during their embryonic development, maternal imprinting acts against the physical growth of the embryo that draws from the host's resources. On the other hand, paternal imprinting tries to promote growth ensuring strong progeny. The situation becomes more complicated by the discovery that certain genes are imprinted in some tissues and not others.

Imprinted genes are often found in clusters in which a common regulatory mechanism is responsible for the expression and repression of the appropriate set of genes in the cluster. A well studied case is that of the *Igf2/H19* locus (Thorvaldsen *et al.* 1998a; Thorvaldsen *et al.* 1998b; Bell and Felsenfeld 2000; Hark *et al.* 2000; Murrell *et al.* 2004; Ling 2006; Engel *et al.* 2006). In this case an enhancer element is located far downstream of the two genes. In the paternally derived chromosome, the differentially methylated region (DMR) that is located between the two genes is methylated, and the enhancer acts on the distal *Igf2* gene inducing its expression. On the maternal chromosome, the DMR is free of methylation and bound by CTCF, which acts as an insulator that cuts off communication of the enhancer with the *Igf2* gene and causes it repression. This in turn allows the more proximal *H19* gene to be expressed. CTCF binding has been reported for the imprinted *Rasgrf1* locus (Yoon *et al.* 2005) and also seems to have a general role on insulator activity (Bell *et al.* 1999; Ohlsson *et al.* 2001).

Another mechanism for the regulation of expression in imprinted clusters is the one found at the *Igf2r/Air* system (Wutz *et al.* 1997; Zwart *et al.* 2001; Sleutels *et al.* 2002; Sleutels *et al.* 2003). In this case methylation of the *Igf2r* promoter and transcription initiation from within the *Igf2r* gene causes expression of the antisense non-coding transcript *Air*. When *Air* is expressed, the other imprinted genes of the locus are repressed. Methylation of the *Air* promoter on the other hand, suppresses its transcription allowing uniparental expression of the imprinted genes. A mechanism that involves a non-coding antisense transcript is present in several known imprinted loci (for an overview see O'Neill 2005) and bears significant similarities with the *Xist/Tsix* system of X chromosome inactivation that will be discussed in more detail in the next section.

1.5.3. Role in X inactivation

Sexual reproduction has an unexpected consequence, the different karyotype between males and females. This means that depending on its gender, an organism will be diploid or aneuploid for one of the sex chromosomes, which could lead to different expression levels of the associated genes with unpredicted effects on the phenotype. Different organisms seem to deal differently with the problem of dosage compensation; in *D. melanogaster* the unique X chromosome in males is doubling its transcription rate, in *C. elegans* both the X chromosomes of hermaphrodites are halving their transcription, while one of the two X chromosomes of female mammals undergo inactivation. The latter mechanism of inactivation is best studied in mammals and is directly associated with chromosome-wide epigenetic modifications (Riggs and Pfeifer 1992).

There are two important decisions to be made during X chromosome inactivation is mammals, if inactivation is needed (*i.e.* counting) and which X chromosome to inactivate (*i.e.* choice). In mouse extraembryonic tissues, it is always the paternal X chromosome that is inactivated, in the embryo proper though the choice seems to be random. Recent research has shown that counting occurs probably through the transient pairing of the two X chromosomes at the *Xite* locus of the X-inactivation centre (*Xic*) (Xu *et al.* 2006; Bacher *et al.* 2006). *Xic* contains three elements, *Xist*, *Tsix* and *Xite* that are all transcriptional units of non-coding RNAs, of which *Xist* and *Tsix* are in opposite orientation, partially overlapping. *Xite*

is responsible for *Tsix* expression in *cis* (Ogawa and Lee 2003). In the chromosome where *Tsix* is expressed, the corresponding *Xist* is suppressed, and the chromosome remains active (*Xa*) (Stavropoulos *et al.* 2001). The opposite is true for the inactive X (*Xi*). In all cases, repression of *Xist* or *Tsix* is associated with DNA methylation at the corresponding promoters.

The *Xist* non-coding RNA is believed to initiate inactivation of the chromosome by coating it (Penny *et al.* 1996; Marahrens *et al.* 1998; Wutz and Jaenisch 2000) but the exact mechanism is not yet fully understood. Heterochromatinisation is then thought to spread across the *Xi* causing chromosome-wide silencing. Importantly, it has long been known that DNA methylation in the *Xi* takes place after the inactivation (Lock *et al.* 1987). There is evidence that complicate the picture as they show that, on average, *Xi* methylation levels are not uniformly higher than in the *Xa* (Viegas-Pequignot *et al.* 1988; Weber *et al.* 2005; Hellman and Chess 2007). It seems that the distribution of DNA methylation differs between the two X chromosomes; *Xa* has high intragenic DNA methylation while *Xi* has high proportion of CpG island and promoter methylation. Further experimental evidence is required to elucidate the X inactivation process.

1.5.4. Role in development

Establishment and maintenance of the DNA methylation patterns are of utmost importance for the correct developmental program of mammals. This is exemplified by the effect that DNA methylation depletion has on mice; deletion of the catalytic domains of DNMT1 (Li *et al.* 1992) or DNMT3a and b (Okano *et al.* 1999) resulted in a dramatic reduction in the genomic methylation levels and embryonic lethality. The same gene deletions however did not seem to affect the growth of embryonic stem cells in culture (Tsumura *et al.* 2006). This establishment of the genome-wide DNA methylation patterns in mice occurs mainly during the two massive waves of epigenetic reprogramming, one in germ cells and one in the early embryo (Figure 1-6). Although this epigenetic reprogramming also involves big changes in the pattern of H3K27 methylation by the polycomb group proteins and fluctuations in the levels of H3K9 and H3K4 methylation (Seki *et al.* 2005), the

DNA methylation changes that are associated with this event are the focus of this section.

Reprogramming of DNA methylation during germ cell differentiation

Epigenetic reprogramming of the germ cells serves mainly to establish all the gender-specific methylation marks independently of the marks in the soma and

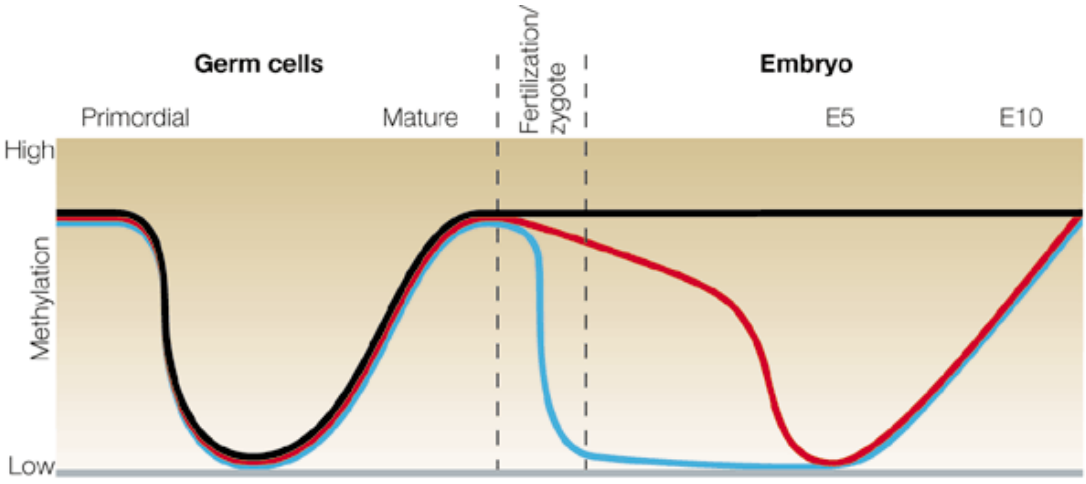

Figure 1-6. Simplified diagram of DNA methylation waves during mouse development. The black line represents imprinted genes, the red line the maternal genome and the blue line the paternal genome. The methylation levels at the y-axis are not to scale. (Taken from Reik and Walter 2001).

happens during maturation of the primordial germ cells (PGC). The PGC are of mesodermal origin and appear as a small population of alkaline phosphatase-positive cells at Embryonic Day (E) 7.0 in mouse soon after the formation of the primitive streak (Ginsburg *et al.* 1990). Subsequently, during gastrulation, this initial population of cells migrates towards the hindgut at the genital ridges and proliferates, processes that last until E11.5 and E13.5 respectively (Hogan *et al.* 1994). The PGC then enter the phase of sex-specific maturation to germ cells, which in males is completed before birth with mitotic arrest and with mitotic arrest at the diplotene stage of meiosis in females. After birth, the events that cause reproductive maturation of the organism trigger the final stage of germ cell maturation into sperm and oocytes.

During the embryonic stage of differentiation (or de-differentiation from another point of view) and despite the presence of DNMT1 in their nucleus (La Salle

et al. 2004), the PGC undergo genome-wide demethylation of DNA. This demethylation involves erasure of imprints, removal of methylation from single-copy genes and repetitive elements and the reactivation of the *Xi* chromosome in females. In more detail, Hajkova *et al.* (2002) have shown a dramatic loss of methylation in imprinted and non-imprinted loci within one day of the end of migration of the PGCsto the genital ridges (E12.5). The authors further speculate on an active mechanism for demethylation, as the extent of methylation loss in the given time frame could not be explained by a replication-dependent passive loss of methylation. Further analysis of different imprinted regions has shown that the exact timing of this reprogramming is individually controlled on each imprinted gene (Lee *et al.* 2002). Contrary to the erasure of imprints, reactivation of the *Xi* appears to be a slow process. It starts as soon as the PGC appear at the primitive streak, continues after the initiation of meiosis and is completed at the mature oocyte (Sugimoto and Abe 2007). The evidence however of how this reactivation is temporally associated with DNA demethylation is scarce. Finally, loss of methylation after E11.5 has also been observed in repetitive elements. The rate of demethylation for these elements is however slow, varies according to the sex of the organism and is not complete (Lees-Murdock *et al.* 2003; Lane *et al.* 2003).

After demethylation, a wave of *de novo* methylation restores methylation on the genome of the mature gametes in a sex-specific manner. In other words the correct imprints are being re-established according to the gender of the organism and the extent of remethylation of the repetitive elements is regulated according to the gamete type too. In more detail, remethylation of repetitive elements such as IAP, L1 and minor satellites in male germ cells is completed by E17.5 (Lees-Murdock *et al.* 2003; Lane *et al.* 2003), while in female germ cells the process begins after birth and is less efficient, resulting in lower overall methylation levels in the oocytes (Walsh *et al.* 1998). Establishment of the imprints also happens asynchronously in the two types of germ cells; in male germ cells methylation starts at around E15.5 and is a slow process that continues after birth (Davis *et al.* 2000; Li *et al.* 2004), while in the female germ cells methylation of the imprinted genes begins post-natally, in the growing oocyte (Lucifero *et al.* 2002; Obata and Kono 2002). Deletion analyses (Bourc his *et al.* 2001; Kaneda *et al.* 2004; Arnaud *et al.* 2006) have shown that

DNMT3a and DNMT3L, both acting together and on their own, are the main methyltransferases responsible for the establishment of the correct imprints and that they are enough for the correct targeting of methylation at this stage. Aditionally, the unique expression of the DNMT3a-β isoform in testes (Weisenberger *et al.* 2002) and the testes- and ovary-specific isoforms of DNMT3L in the corresponding tissues (Shovlin *et al.* 2007) might provide the cell-type specificity needed for correct imprint establishment.

Finally, in contrast to the orderly model of epigenetic reprogramming in the gametes described above, recent evidence has come to light that somehow complicates the phenomenon. Oakes *et al.* (2007) have examined the methylation status of hundreds of genomic loci during distinct stages of post-natal male germ cell maturation. In this study, except from the loci that gained methylation, as expected, several loci seemed to be going against the massive global methylation wave and become demethylated instead. Better appreciation of the mechanisms that underlie the correct methylation and demethylation of specific loci is probably required for the understanding of this biological process.

Post-fertilisation reprogramming of DNA methylation

Soon after fertilisation, the sperm pronucleus decondenses and the DNA becomes demethylated (Mayer *et al.* 2000; Oswald *et al.* 2000). This process is completed before the first division of the zygote, which excludes the possibility of a passive demethylation mechanism and has caused speculations about an active DNA demethylase enzyme. Despite intensive research however the speculative DNA demethylase remains uncharacterised and there is some discussion that this role might be taken temporarily by a DNA repair enzyme (Morgan *et al.* 2004; Gehring *et al.* 2006; Barreto *et al.* 2007). The only elements in the paternal genome that seem to escape active demethylation are the imprinted genes and IAP elements. (Nakamura *et al.* 2007).

The maternal genome also undergoes global demethylation with the exception of the imprinted genes but this process lasts much longer, until the blastocyst stage, and is probably a passive mechanism. The passive demethylation of the maternal genome can be explained by the active exclusion of the oocyte-specific DNMT1 isoform (DNMT1o) from the nucleus during this period (Mertineit *et al.*

1998; Howell *et al.* 2001). There is no explanation as to why there is a need for two separate mechanisms for the demethylation of the paternal and maternal genomes or indeed what purpose does this wave of demethylation serve. The somatic methylation pattern is nevertheless reastablished around implantation. This is also the time of the onset of X inactivation in the embryo proper.

Epigenetic reprogramming in development across organisms

The evidence outlined above comes mainly from experiments in mice. For ethical reasons human embryonic development is not as well studied, but the limited experimental evidence that is available support the existence of similar reprogramming waves during human development as in mouse. Comparative studies of the post-fertilisation reprogramming waves have shown that the mechanism is not conserved in all mammals (Dean *et al.* 2001; Beaujean *et al.* 2004). In these studies, immunostaining of methylcytosine in the male and female pronuclei of the zygote has demonstrated that the male pronucleus of rat and pig, but not that of sheep and rabbit, undergoes active demethylation. The male bovine pronucleus is only partially demethylated before the first zygotic division but does undergo remethylation before implantation.

In the non-mammalian species studied, reprogramming waves seem to be missing. In more detail, studies of the methylation in single-copy genes and repetitive elements in zebrafish at various developmental stages revealed no loss of methylation in the oocytes, sperm or the early embryo (Macleod *et al.* 1999). Similarly, the *Xenopus* zygote does not seem to undergo demethylation and on the contrary appears to depend on DNA methylation for its correct development (Stancheva and Meehan 2000). It looks like the significance of the DNA methylation patterns in fate determination and developmental competence varies among organisms. This situation is reminiscent of the fact that DNA methylation is absent from many invertebrates and begs for an answer to the question of what is the particular contribution of DNA methylation that benefits the organisms that have it.

DNA methylation in ES cells

Embryonic stem (ES) cells are derived from the inner cell mass of the preimplantation blastocyst. As discussed previously, the time of implantation is also

the time when the second *de novo* methylation wave occurs. As a consecuence of that, a mixture of DNA methylation marks, characteristic of both the pre- and post-implantation stages, are present in ES cells. On one hand, the global DNA methylation levels, as well as the methylation levels of repetitive elements, are high in ES cells, like in the epiblast (Jackson *et al.* 2004; Yeo *et al.* 2007). On the other hand, in female ES cells, X inactivation has not yet occured. Additionally, ES cells also exhibit DNA methylation characteristics that are not reflecting the *in vivo* situation; analysis of the methylation status of certain imprinted genes in mouse ES cells showed variation in their methylation levels among different ES subclones (Humpherys *et al.* 2001). It is also important to note that cell lines have been known to accumulate aberrant methylation in tissue culture (Antequera *et al.* 1990) and the possibility that this also happens in ES cells cannot be excluded. The empirical observation that early passage ES cells are more competent for the formation of embryos could be linked to this. Finally, upon induction of ES cell differentiation *in vitro*, there is a reproducible change of the DNA methylation levels in specific loci (Shiota *et al.* 2002; Kremenskoy *et al.* 2003). This is presumed to be linked to the transcriptional regulation of specific genes that are expressed in different cell lineages and is believed to reflect similar changes that happen during development.

1.5.5. Defence against parasitic sequences

The idea that methylation is a defence mechanism against parasitic sequences comes from bacteria where DNA methylation was first detected as a specific mechanism for the protection against invading DNA. In mammals, this view is supported by the observation that in transgenesis experiments, the foreign DNA very often becomes heavily methylated in the host cell. Concrete evidence however for a role of DNA methylation in mammals as a defence mechanism comes from observations on transposons.

Repetitive sequences comprise 40-50 % of the mouse and human genomes respectively (Mouse Genome Sequencing Consortium 2002; Human Genome Sequencing International Consortium 2004), the large majority of which is repetitive elements. Although there is arguably a function for the presence of repetitive sequences in the genome (Shapiro and Von Sternberg 2005) random retrotransposition can have deleterious effects for the organism (Gilbert *et al.* 2002;

Symer *et al.* 2002). It is believed that the high levels of methylation that are found at retrotransposons (Yoder *et al.* 1997) help relieve the load of mutagenesis through transposition in many organisms. Indeed, mice with reduced global methylation levels have been shown to have increased activation of IAP elements (Walsh *et al.* 1998) and methylation seems to inactivate human endogenous retroviral sequences (Lavie *et al.* 2005). Similar studies in *A. thaliana* have confirmed a similar mechanism (Kato *et al.* 2003).

A specialised case of DNA methylation happens is *Neurospora crassa*. This organism has very low levels of DNA methylation that seem to be acting uniquely through the mechanism of Repeat Induced Point Mutation (RIP). In more detail, *N. crassa* seems to inactivate repetitive elements by inducing C to T transitions in their sequence rendering them inactive. The relics of repetitive elements found in this organism's genome are also the only sites where methylation can be found (Selker *et al.* 2003; Freitag *et al.* 2002). Although the exact mechanism of RIP is not yet fully understood, it should be noted that the presence of methylcytosines coincides with regions of C to T transitions. It is possible that in this case *N. crassa* might have taken advantage of methylcytosine deamination for its benefit (Selker 1990). This is also supported by the observation that the *rid-1* (*RIP- defective*) gene, which is required for RIP, encodes a putative DNA methyltransferase protein (Freitag *et al.* 2002).

1.5.6. Role in chromatin structure and integrity

Nucleosomes comprise the most elementary level of DNA compaction, in which the DNA is wrapped around the core histone octamer. It has been known for a long time that certain DNA sequences make better nucleosome positioning signals than others. *In vitro* experiments on the strong nucleosome positioning sequence of the chicken β-globin promoter have shown that the presence of certain central CpGs is essential for nucleosome positioning (Davey *et al.* 2004) and that their methylation could displace the *in vitro* reconstituted nucleosome (Davey *et al.* 1997). A similar effect of DNA methylation on some, but not all, nucleosome positions was detected in the human H19 imprinting control region (Davey *et al.* 2003). However, all these experiments are on reconstituted chromatin, it remains to be shown whether the

effect that DNA methylation has in nucleosome assembly and positioning has a biological meaning.

The next level of chromatin compaction is the 30 nm fibre that spontaneously forms *in vitro* by the addition of histone H1. The experimental evidence for a causative role of methylcytosines in H1 recruitment is inconclusive. Different reconstitution and electrophoretic mobility experiments of methylated and unmethylated DNA sequences have produced conflicting results that show both a preference and indifference of H1 for methylated DNA (Nightingale and Wolffe 1995; Campoy *et al.* 1995; McArthur and Thomas 1996). An older *in vivo* study however that employed differential sedimentation of mono- and oligo-nucleosomes and immunodetection of methylcytosines in the purified fractions had shown a clear preference of H1-containing dinucleosomes for methylated DNA (Ball *et al.* 1983). This experiment however does not exclude the possibility that there is a third factor who brings the two together and not DNA methylation *per se*. Moreover, fluorescence recovery after photobleaching (FRAP) experiments on DNMT3a/b-depleted and wild-type ES cells revealed reduced motility of H1 in the former (Gilbert *et al.* 2007). The same authors however showed that the absence of DNA methylation from the genome of the mutant cells did not affect general chromatin compaction. On the other hand, immunoprecipitation experiments showed no preference of H1 for a methylated or an unmethylated version of a transgene (Hashimshony *et al.* 2003). It looks like more experiments are needed to show whether DNA methylation recruits H1 and what its role on heterochromatin formation is.

A well studied function of DNA methylation on chromatin structure is that at nucleolus organising regions (NORs). NORs contain the portion of the genome that encodes for the 18S, 5.8S and 25S ribosomal RNAs (rRNA). These genes are found in tandem repeats that are separated by a spacer and can span a region of several megabases. Active NORs form secondary constrictions in metaphase chromosomes and make up the structures that are recognised as nucleoli during interphase. A NOR is active when transcribed by RNA polymerase I and inactive when transcription is inhibited by methylation at the spacer region (Bird *et al.* 1981; Labhart 1994). In this way, DNA methylation –or better the absence of it– plays a role on nucleoli and

secondary constriction formation. Recent research has shed light on the mechanism of NOR activation which appears to be regulated by specific transcription factor binding and heterochromatinisation, similarly to the activation of polymerase II transcribed genes (Laengst *et al.* 1998; Zhou *et al.* 2002; Santoro *et al.* 2002). Further detailed analysis of this mechanism however, is beyond the scope of this introduction.

Perhaps the best evidence that DNA methylation has an active role on chromatin structure and integrity comes from the fact the immunodeficiency, centromeric instability and facial anomalies (ICF) syndrome in humans is caused by mutations in DNMT3b (Hansen *et al.* 1999). As its name implies, some of the phenotypes of this syndrome are decondensation of the centromeres of chromosomes 1, 9 and 16, chromosomal rearrangements, and reduced methylation of satellite DNA repeats. Importantly, most of the known DNMT3b mutations in patients with ICF affect the catalytic function of the protein and not protein-protein interactions (Lappalainen and Vihinen 2002). Further evidence for a structural role of DNA methylation at the centromeres comes from the observation that all three DNMT1, DNMT3a and DNMT3b seem to be preferentially localised at the centromeres of both human and mouse chromosomes during metaphase (Craig *et al.* 2003). The role of DNA methylation on centromere formation was shown more directly in experiments where demethylation by 5-azacytidine caused redistribution of the central centromeric scaffold protein CENP-B (Mitchell *et al.* 1996). All of this evidence supports the idea that DNA methylation at centromeres is not merely a consequence of their heterochromatinisation, but plays an active role in their formation.

1.6. Perspective

As shown in the previous pages, there is a wealth of information about the molecular machinery that is associated with the DNA methylation patterns. Moreover, we have a relatively good account of when and where DNA methylation

occurs, although there is definitely room for more descriptive studies of this type. Our understanding of the DNA methylation phenomenon however does not seem to have leaped accordingly. On the contrary, it seems like the more information comes to light about all the phenomena that surround DNA methylation, the less we understand it. It is for example now known that DNA methylation is not universal among eukaryotes and that different organisms have different patterns and levels of methylcytosine in their genomes. There are many theories which try to explain these patterns, but with the exception of a few cases, a definitive answer to why and how they exist is still beyond our grasp. Moreover, a link between DNA methylation and transcriptional repression has been known for a long time and it seems that many of the roles of methylation are based on this link. Recent evidence however indicates that DNA methylation might actually follow transcription repression and that the universally conserved histone modifications are temporally more closely related to transcription. If this is the case, why should methylation be established at already inert promoters? The question becomes more urgent by the realisation that methylation establishment only happens in a subgroup of organisms. Which evolutionary process and biological function would necessitate this DNA modification in some organisms and not others?

These questions have been the motivation behind the research work presented here. In the following chapters, the significance of DNA methylation for transcriptional repression and correct embryonic development as well as the formation of the methylation patterns in the mammalian genome are explored. In more detail, in the third chapter the process of silencing of the *Oct4* gene during development is used as a model for the investigation of how methylation is established and what it contributes to the silencing process. A genome-wide analysis of the distribution and the methylation status of CpG islands in mouse and the possible existence for a developmental program for the establishment of the CpG island methylation is presented in the fourth chapter. An overview of the results with regard to the questions posed here is presented at the end.

2. Materials and methods

2.1. Murine embryonic stem cell culture and differentiation

2.1.1. Cell lines

The mouse embryonic stem (ES) cell lines used in this study were the wt E14 (129/Ola background) (Handyside *et al.* 1989) , DNMT1 KO (129/Ola background) (Li *et al.* 1992), DNTM3a/b KO 7aabb (129/Ola background) (Okano *et al.* 1999 and wt and G9a-/- derivatives of the TT2 ES cell line (C57BL/6 background) (Yagi *et al.* 1993), COL4 and 2-3 (Tachibana *et al.* 2002) respectively. The DNMT1 KO and 7aabb cell lines were a gift from the lab of Dr Bernard Ramsahoye. The COL4 and 2-3 cell lines were a gift from the lab of Dr Amanda Fisher.

2.1.2. Mouse embryonic stem (ES) cell tissue culture

The stem cells were grown on precoated gelatinized flasks at 37^{o}C in the presence of 5% CO. The full medium was 1x Glasgow modified Eagle's Medium (GMEM) (Invitrogen) supplemented with 1 mM sodium pyruvate (Invitrogen), 1x non-essential amino acids (Invitrogen), 10% foetal bovine serum (HyClone) previously tested for stem cell culture suitability, human recombinant leukaemia inhibitory factor (LIF) and 1:1000 β-mercaptoethanol (Invitrogen). The cells were dissociated by incubating briefly with 0.05% v/v trypsin (Cambrex) in PBS, prewarmed at 37^{o}C The trypsin was inactivated with an equal volume of full medium and the cells were pelleted by spinning at 1300 rpm for 3 min. The cells were split approximately every two days to a confluency of 30%.

2.1.3. In vitro differentiation of mouse ES cells to embryoid bodies

The embryonic stem cells were trypsinized and pelleted and then washed with full medium without LIF. One T75 flask of confluent cells was transferred to a 100 mm^2 bacteriological Petri dish with 15 ml full medium without LIF and incubated at 37^{o}C in of 5% CO. From this point on, all handling of the cells differentiating in suspension was done with a 25 ml pipette with wide orifice to avoid disaggregation of the cells. The medium was changed the next day and then every two days for the

desired differentiation period. The embryoid bodies were harvested after 3, 7, 14 and 21 days in culture without LIF (EB3, EB7, EB14 and EB21 respectively). Alternatively, 1:10,000 of RA stock solution was added to the full medium without LIF after the EB3 stage and the cells were harvested after 2, 4, 6 or 10 days (RA2, RA4, RA6, and RA10 respectively). For harvesting, the embryoid bodies were transferred to a 30 ml centrifuge tube and allowed to settle at the bottom of the tube, the medium was aspirated and the cells washed with PBS twice.

2.2. Isolation of high molecular weight genomic DNA

Two separate protocols were used for genomic DNA extraction, one for cells grown in culture (ES cells and embryoid bodies) and one for the extraction of DNA from mouse brains. The protocol used for DNA extraction from cells grown in culture relies on the denaturation of most proteins under very high salt conditions. A phenol/chlorophorm step was nevertheless added at the end, to ensure the absence of nucleases from the sample for long-term storage. For the extraction of DNA from brain, the emphasis lies on efficiently lysing the tissue sample and has a very thorough phenol/chlorophorm extraction step at the end to remove all proteins. In either protocol the concentration of the isolated DNA was determined with a NanoDrop ND-1000 spectrophotometre. The presence of contaminant proteins was assessed by the A_{260}/A_{280} ratio which needed to be close to 2 and above 1.8. The DNA quality was confirmed with 0.8% agarose gel electrophoresis in 1x TAE buffer with 0.5 µg/ml ethidium bromide.

2.2.1. DNA extraction from cells grown in tissue culture

2 ml of Lysis buffer I was used for approximately $5x10^6$ cells, precipitated and washed in PBS as described before. The cells were allowed to lyse overnight at 37^o C with mild agitation. The next day, 1/3 volume of saturated NaCl was added to the lysate and the mix was shaken vigorously for 30 sec. At this stage a lot of protein

foam should appear. Next, the sample was centrifuged at room temperature at 900g for 30 min. The supernatant was transferred to a new tube and precipitated with two volumes of ice-cold absolute ethanol. The DNA was recovered with a Pasteur pipette hook, washed with 70% ethanol and resuspended in 1-2 ml TE buffer. A final concentration of 0.1 mg/ml RNase A was added and the sample was left at 37^{o} C for two hours with mild agitation. The DNA was extracted once with an equal volume of SAGE solution, precipitated and resuspended in TE buffer.

2.2.2. DNA extraction from mouse brain

The mice that were sacrificed were all wild type, between four and six months old from a C57BL/6 and 129/Ola mixed background. The brains were dissected from the skull, sectioned free of the spinal cord and optical nerve and washed in PBS. Brains of three animals of the same sex were pooled before DNA extraction. The pooled brains were transferred into approximately five volumes of Lysis buffer II and the tissue was homogenised with a Dounce homogeniser. A final concentration of 1% w/v SDS and 0.4 mg/ml Proteinase K was added and the brains were lysed over night at 55^{o} C. Next day, RNase A was added to a final concentration of 0.1 mg/ml and the RNA was digested at 37^{o} C for 2 hours. The DNA was extracted with SAGE solution for 2-4 times and the DNA precipitated with two volumes ice-cold ethanol, recovered with a Pasteur pipette hook and resuspended in TE buffer.

2.3. Bisulfite genomic sequencing

Sodium bisulfite causes the specific deamination of cytosine through a sulfonated intermediate, and its conversion to uracil. The bisulfite genomic sequencing principle relies on the fact that the reaction rate for 5' methylcytocine is much slower than that of cytocine Wang *et al.* 1980 making it virtually impossible to convert methylated cytocines to uracils by sodium bisulfite. The treated DNA is then PCR amplified and sequenced. Comparison of the sequenced DNA with the original

genomic sequence reveals which cytosines were resistant to the treatment and therefore methylated.

2.3.1. Bisulfite treatment.

The protocol used for the bisulfite treatment is modified from Frommer *et al.* 1992. In more detail, 2µg of genomic DNA were digested at 37^{o} C for 8 hours with 10U of the restriction endonuclease Kpn I (New England Biolabs) in a 50 µl reaction according to the manufacturer's instructions. The digestion facilitates the complete denaturation of DNA at the next stage. The digested DNA was extracted with SAGE solution, precipitated with absolute ethanol and resuspended in 25 µl TE buffer. The DNA was then boiled for 5 min and further denatured for 20 min in 42^{o} C in the presence of 0.3 M NaOH. This stage is very important as non-denatured, double stranded DNA can not be converted by sodium bisulfite. In the next step, the denatured DNA was incubated in a total volume of 300 µl with freshly prepared sodium bisulfite solution. The reaction proceeded under mineral oil at 55^{o} C in the dark. After exactly 5 hours, the DNA was precipitated and resuspended in 25 µl TE buffer. Next, the treated DNA was desulfonated by addition of 2.5 µl 3M NaOH and incubation at 37^{o} C for 15 min. Finally, the DNA was once more precipitated by adding 32.5 µl 5 M ammonium acetate (pH 7.0) and 180 µl absolute ethanol and then washed with 300 µl 70% ethanol and resuspended in 25 µl TE buffer.

2.3.2. PCR amplification and sequencing

The bisulfite-treated DNA was PCR amplified with primers specific for the genomic region of interest. The primers were designed so that they were not interrupted by a Kpn I restriction site and in some cases nested PCR was performed in order to acquire a specific product. A list of the primers used for bisulfite sequencing in this study, their annealing temperatures and product size is given in Table 2-1. The PCR program was: 92^{o} C 1min, 35x(92^{o} C 30 sec, T_{an} 30 sec,72^{o} C 30 sec), 72^{o} C 5 min, on 2 µl of the treated template. The PCR reactions were performed in 30 µl with 0.5 U Fast Start Taq DNA polymerase (Roche), 0.2 mM dNTPs (ABgene) 0.6 µM of each primer in the reaction buffer provided by the manufacturer (contains 2 mM $MgCl_2$). The amplified product was resolved in a 2% agarose gel in 1x TAE buffer, with 0.5 µg /ml ethidium bromide. The band of the

Table 2-1. Primers used in bisulfite PCR.

Primer	Sequence	Primer	Sequence	bp	T_{an}
oct4 fw (-208)	TTTGAAGGTTGAAAATGAAGTTTT	oct rev (+106)	CATCACCCCCACTAATAAAAATAA	386	63° C
as above (nested)		oct rev (+55)	CAACCATAAAAAAAATAAACACCCC	263	63° C
oct fw (-485)	GTTGTTTTGTTTTGGTTTTGGATAT	oct rev (-235)	AATCCTCTCACCCCTACCTTAAAT	250	58° C
oct fw (-848)	AGGTTTTTTTG ATTTGAAGTAGA	oct rev (-535)	AACTCTACACCATAAAACCCC	313	60° C
oct fw (-1199)	AGGGTAGGTTT TTGTATTTTTTTT	oct rev (-983)	ACTCCCCTAAAAACAACTTCCTACT	216	60° C
as above (nested)		oct rev (-1027)	5'-CAATCCCCTCA CACAAAAC-3'	172	59° C
oct fw (-1670)	GTGTTATGTGTAGTTGTGTGTAGGT	oct rev (-1341)	TTATCTATCTACTCCTACACCATACT	329	60° C
oct fw (-2088)	GGTTTTAGAGGTTGGTTTTGGG	oct rev (-1749)	CATCTCTCTAACCCTCTCCATAAATC	339	63° C
oct fw (-2070)	TGGGAGGAATTGGGTGTG	as above (nested)		321	63° C
celsr2 bis ex1 fw	AGTTTTTTAGAGATTTTTATTAGGGT	celsr2 bis ex1 rev	AACTAATAACATACCCTTATCCACC	263	56° C
celsr2 bis ex2 fw	TTTGTTTTTAGGGATTTTTAGGAG'	celsr2 bis ex2 rev	CAAAAACAAATATAACCACCCTCTC	344	61° C
as above (nested)		celsr2 bis ex2-2 rev	ATCTTAAATAACAAAAATCCACCTC	290	50° C
celsr3 ex1 fw	GAAGAAGGTTATTTTATTTTGTAGTT	celsr3 ex1 rev	AACTTTCCATAATTCCCTATCCAC	187	56° C
celsr1 (3-1) fw	GTGTTAGTGTTTGAGAATGAGTTTG	celsr1 (3-1) rev	AATACATAAATATCCTTAATCTCCC	209	54° C
celsr1 (3-2) fw	GGTGGGTTATGAGGTTTTGATTAT	celsr1 (3-2) rev	TCATTAACCTCACCAACAACTAA	192	60° C
celsr1 (4-2) fw	GTTTGAGGTTA TTATTAATTTTTT	celsr1 (4-2) rev	CAACTCAACTAAACCTCCTAACC	106	50° C
T7 fw	GTAATACGACTCACTATAGGGC	M13 rev	GTAAAACGACGGCCAG		54° C

correct size was excised from the gel and the DNA was extracted using the QIAquick Gel Extraction kit (Qiagen).

The amplified DNA was cloned into the TOPO-TA vector (Invitrogen) according to the manufacturer's instructions. Sufficient number of white colonies was picked with a sterile pipette tip the next day and the insert was amplified directly from the colonies using the T7 forward and M13 reverse primers (Table 2-1). These primers are complementary to the vector sequence and flank the insertion site adding 180 bp to the amplification product. The colony-PCRs were performed in 20 µl with 0.5 U Red Hot DNA Polymerase (ABgene), 1.5 mM $MgCl_2$, 0.25 mM dNTPs (ABgene) and 0.2 µM of each primer, in the reaction buffer of the manufacturer. The PCR program was: 95° C 1min, 30x(95° C 1 min, 54° C 1 min,72° C 1 min), 72° C 7 min. The correct size of the PCR product was confirmed with agarose gel electrophoresis and 15 µl of each PCR reaction were treated at 37° C for 30 min with 5 U Exonuclease I (New England Biolabs) and 0.25 U Shrimp Alkaline Phosphatase (Roche), followed by inactivation of the enzymes at 80° C for 15 min. This reaction is important for removing any unused primers and dephosphorylating unused dNTPs before the sequencing reaction.

The sequencing reaction was performed in 10 µl reactions on 3ul of the treated PCR product with 2 µl BigDye Terminator v3.1 (Applied Biosystems) reaction mix, 8 mM Tris-HCl pH8, 0.25 mM $MgCl_2$ and 0.16 µM of either the M13 reverse or T7 forward primers. The sequencing was performed on an ABI 3730 sequencer.

2.3.3. Analysis of the sequencing data

Analysis of the sequenced results was performed with the aid of the software BiQ Analyser (http://biq-analyzer.bioinf.mpi-sb.mpg.de). In more detail, the original genomic sequence was aligned with the sequenced clones and the quality of the sequences was assessed. The efficiency of the bisulfite conversion was judged by the absence of non-converted cytosines in a non-CpG context in the sequencing result and clones with conversion rates below 90% were removed. Similarly, clones that shared homology with the genomic sequence below 80% or had clonal methylation patterns were all removed. The only exception was the specific case of homogeneously methylated or homogeneously unmethylated clones which were

included in the results. The methylation pattern of all the clones that had passed the quality control was then recorded and graphically represented.

The statistical analysis of the bisulfite data was performed as follows. For the comparison of the methylation patterns in the various 5' upstream regions of the *Oct4* gene the Kolmogorov-Smirnov and the Wilcoxon two sample rank tests were performed. Both tests are non-parametric and do not make a normal distribution assumption. The Kolmogorov-Smirnov test investigates against the null hypothesis that the data of the two samples come from the same distribution, while the Wilcoxon test has the null hypothesis that the medians of the two samples do not vary significantly. The null hypotheses were rejected if $P \leq 0.05$. The values that were used in the statistical analysis of each sample were the proportion of methylated versus total CpGs per clone sequenced for each region, thus the sample size was the number of clones sequenced in each case.

2.4. Analysis of RNA

All the plasticware and all the solutions used in RNA analysis were treated with DEPC.

2.4.1. RNA isolation

RNA was isolated with the TRI reagent (Sigma) according to the manufacturer's instructions. 1 ml TRI reagent was used for every 10^7 cells. Great care was taken to avoid contact with the DNA-containing interphase during the RNA extraction. The RNA was redissolved in RNase and Nuclease-free water (Ambion) containing 1:100 RNasin RNA inhibitor (Promega). The concentration of the RNA was determined with a NanoDrop ND-1000 spectrophotometre. The presence of contaminant proteins was assessed by the A_{260}/A_{280} ratio which needed to be close to 2 and above 1.8.

The RNA quality and quantity was confirmed with denaturing gel electrophoresis. In detail, 1 μg RNA, as measured, was heated in 1x RNA loading buffer at 70^{o} C for 5 min and then transferred to ice. The RNA was resolved in a

1.2% agarose gel in 1x MOPS buffer with 0.24 M formaldehyde and 0.5 µg /ml ethidium bromide.

2.4.2. Reverse transcription

2 µg RNA were reverse transcribed in a 20 µl reaction containing 1.5 mM dNTPs (ABgene), 1x hexanucleotide mix (Roche), 1 µl of RNasin ribonuclease inhibitor (Promega), 1x first-strand buffer (Invitrogen), 10 mM DTT (Invitrogen) and 200 U M-MLV Reverse Transcriptase (Invitrogen). The RNA was first mixed with the random hexanucleotides and dNTPs, denatured at 70^{o} C for 10 min and then quickly cooled in ice. The reaction was incubated at 37o C for 1 h and the enzyme was inactivated at 70o C for 15 min. For each RNA sample, a mock reverse transcription reaction was performed that contained 1 µl of water instead of reverse transcriptase. Typically, a 1:20 dilution of the reaction was used in PCR.

2.4.3. RT-PCR

The primers used for RT-PCR, their sequences, annealing temperatures and product length are listed in Table 2-2. The reactions were carried out in 15 µl volume with 1.5 U Taq DNA polymerase (Roche), 0.25 mM dNTPs (ABgene), 0.2 µM of each primer, in the reaction buffer provided by the manufacturer (contains1.5 mM $MgCl_2$). *Gapdh* was used as a loading control. The elongation times were 20 sec for the shorter products and 40 sec for the longer ones. The reactions were not allowed to proceed for more than 35 cycles with most reactions repeated for 30 cycles.

2.4.4. Quantitative real-time RT-PCR (RT-qPCR)

The primers used for RT-qPCR are listed in Table 2-3. All the primers have efficiency between 80 and 120%. The efficiency of the primers was calculated on 10-fold dilutions of cDNA using the iCycler software (BIO-RAD). All the primers were designed so that they produced a product 70-130 nt long, so that the PCR reactions could be performed with a simple two-step program: 95^{o} C 2 min, 40x(95^{o} C 30 sec, 65^{o} C 45 sec). All the reactions were done in quadruplicates. The specific amplification of a single product was ensured by confirming that a single melting curve peak was produced after each reaction. Since the *gapdh* primers could also amplify genomic DNA, control reactions with mock-reverse transcribed RNA were

Table 2-2. Primers used in RT-PCR

Primer	Sequence	Primer	Sequence	bp	T_{an}
Gapdh1 fw	GACTTCAACAGCAACTCCAC	Gapdh1 rev	TCCACCACCCTGTTGCTGTA	125	65° C
Hnf4a 1	ACACGTCCCCATCTGAAG	Hnf4a 2	CTTCCTTCTTCATGCCAG	269	56° C
Sox17 fw	CGAGCCAAAGCGGAGTCTC	Sox17 rev	TGCCAAGGTCAACGCCTTC	154	58° C
Alb 1	GATGAAACATATGTCCCCAAAGA	Alb 2	TGTGTTCCTAGGGTGTTGATTTTA	500	60° C
a-FETO fw	GGAGGCTATGCATCACCAGT	a-FETO rev	CCGAGAAATCTGCAGTGACA	206	55° C
TTR 1	AGTCCTGGATGCTGTCCGAG	TTR 2	TTCCTGAGCTGCTAACACGG	440	58° C
TRA 1	TCCTGCTGATTCCGAATG	TRA 2	TGGCACAGGAACACTTTG	178	58° C
Pax-6 1	AGACTTTAACCAAGGGCGGT	Pax-6 2	TAGCCAGGTTGCGAAGAACT	561	64° C
Msx fw	ACTCCCCTTCAGCGTCGAGTCTCTG	Msx rev	GCGTCCGTGGTTTGCGATTG	202	68° C
Sox1 fw	TGCAGGAGGCACAGCTGGCCTAC	Sox1 rev	TGCCGCCACCGCCGAGTTCTGG	171	57° C
Fgf5 fw	GCGACGTTTTCTTCGTCTTC	Fgf5 rev	ACAATCCCCTGAGACACAGC	239	64° C
Flk-1 1	CCTGGTCAAACAGCTCATCA	Flk-1 2	AAGCGTCTGCCTCAATCACT	599	60° C
Nkx 1	ACATTTTACCCGGGAGCCTACGGTG	Nkx 2	CTTTCCGTCGCCGCCGTGCGCGTG	152	60° C
Tpbpa 1	AATCTTCCTAGTCATCCTATGCC	Tpbpa 2	CGCCACTCTCTGTGTAATCC	331	59° C
REX1 fw	GGACTAAGAGCTGGGACACG	REX1 rev	CCTGCTTTTTGGTCAGTGGT	161	61° C

performed for them to ensure that this was not the case. All the other primers were either interrupted by introns or spanned very large introns. Finally, all the reactions had no-template controls to ensure there was no cross-contamination between wells. The PCR reactions were performed in 25 µl, on 2.5 µl of the 1:20 diluted cDNA, 1x SYBRGreen master-mix (BIO-RAD) and 0.2 µM of each primer. The quantity of each RNA species was determined relative to *gapdh* by using the formula:

$$2^{-(\dot{C}_{T(gene)} - \dot{C}_{T(gapdh)})}$$

where $\dot{C}T_{(gene)}$ is the mean threshold cycle of the particular gene under examination and $\dot{C}T_{(gapdh)}$ the mean threshold cycle of *gapdh*.

Table 2-3. Primers used in RT-qPCR

Primer	Sequence	Primer	Sequence
Gapdh3	TACCCCCAATGTGTCCGTCG	Gapdh4	CCTGCTTCAGCACCTTCTG
Oct RT fw	GAGGAGTCCCAGGACATGAAAGC	Oct RT rev	CCTTTCCAAAGAGAACGCCCAGG
Nanog fw	TGGGAACGCCTCATCAATGC	Nanog rev	AGGTCTTCAGAGGAAGGGCG
D3a-RT-F-Ex6	GTGTCTTGGTGGATGACAGGC	D3a-RT-R-Ex7	GCGGCATGAGCTTCTCCACAC
D3b-RT-Ex7_F	TGGTGTCCTGGAAAGCCACCT	D3b-RT-Ex8-R	AGCCACCAGTTTGTCAGCAGA

2.4.5. Northern hybridization

The radio-labelled probes used for the Northerns were prepared as follows; the region of interest was PCR-amplified from cDNA and the PCR product was gel-extracted from a 0.8% agarose gel in 1x TAE using the QIAquick Gel Extraction kit (Qiagen). The *Oct4*-specific primer sequence was 5'-GTGGTTCGAGTATGGTTC-3' and 5'-AATGATGAGTGACAGACAGG-3' and the amplified region was a 452 bp-long fragment of the gene's last exon. The probe used as a loading control was against the S26 RNA and was a kind gift of Dr J. Selfridge. 3 µl of the gel-extracted probe were then diluted with 6 µl water and heated at 100^{o} C for 10 min. The reaction mix was prepared with addition of 1x hexanucleotide mix (Roche), 0.5 mM of each dATP,dTTP and dGTP, 5 µl of newly ordered $dCTP^{32}$ and 5 U Klenow DNA polymerase I (New England Biolabs) and incubated at 37^{o} C over night. Any unincorporated $dCTP^{32}$ was removed with a Sephadex G-50 NICK Column (Amersham) according to the manufacturer's instructions.

10 µg RNA were resolved in a 1.2 % denaturing agarose gel as described previously (2.4.1). The gel was washed with 10x SSC for 15 min and the RNA was transferred to a Hybond-N+ membrane over night with capillary blotting, using 10x SSC buffer. After the transfer, the RNA was cross-linked on the wet membrane by

UV irradiation (70mJ/cm^2). The membrane was then blocked for 30 min at 65^o C with prewarmed modified Church and Gilbert buffer. The buffer was then replaced by 0.2 ml/cm^2 modified Church and Gilbert buffer containing the radiolabelled probe and the hybridization continued at 65^o C overnight.

The washes were performed in room temperature as follows: 2x low stringency wash (2xSSC, 0.1% w/v SDS) for 5 min, 2x medium stringency wash (1xSSC, 0.1% w/v SDS) for 10 min and 1x high stringency wash (0.1xSSC, 0.1% w/v SDS) for 5 min. The blot was exposed to a phosphor screen (Amersham) for 2 h and the signal was scanned with a STORM Phosphorimager (Amersham). The image analysis and signal quantification was performed with the Scion Image software.

2.5. Chromatin immunoprecipitation (ChIP)

2.5.1. Sample preparation

10^8 cells (one T75 flask) were harvested and resuspended in 10 ml medium with 1% v/v formaldehyde and incubated in the rocking platform for 5 min. The crosslinking was quenched with addition of glycine to a final concentration of 125 mM for 5 min. The cells were pelleted at 1.5 rpm and washed with 20 ml PBS containing protease inhibitor cocktail (Roche). The crosslinked cells were resuspended in 1 ml NE1 buffer and disrupted with 5-10 strokes in a Dounce homogeniser. The nuclei in the suspension were pelleted at 3000rpm for 5 min and resuspended 500 µl NE1 buffer. At this stage the cells were frozen at -80^o C if needed.

The prepared chromatin was mildly digested with 50 U micrococcal nuclease (Fermentas) in ice for 2 h in the presence of 2 mM CaCl and the reaction was stopped with 100 mM EDTA. The samples were then sonicated with a digital Branson sonifier S-450D for 10 min (30 sec on/45 sec off) in an ethanol-ice bath to avoid sample overheating. The final average size of the DNA was 400 bp.

2.5.2. Immunoprecipitation

200 µl of sonicated chromatin were pre-cleared with 60 µl protein A beads (Amersham) at 4° C for 1h. The beads had been previously blocked with 5 mg/ml tRNA and 0.1 mg/ml BSA for 2 h and equilibrated with NE1 buffer. The beads were removed with a brief spin and the chromatin was incubated with rotation with 10 µl antibody (or nothing for the no antibody control) over night at 4° C. The antibodies used in this study are shown in Table 2-4. Next day 60 µl of protein A beads prepared as above were added and the incubation continued for 3.5 h.

Table 2-4. Antibodies used in ChIP.

Company	Affinity	Catalogue no
Upstate	acH4K(5,8,12,16)	06-598
Upstate	acH3K(9,14)	06-599
Upstate	3meH3K4	07-473
Upstate	2meH3K9	07-441
Upstate	3meH3K9	07-442
Upstate	3meH3K27	05-851
Upstate	3meH4K20	07-463

The beads were collected by spinning the sample briefly and washed with 1 ml of the following buffers for 10 min: once with TSE I, four times with TSE II, once with Buffer III and thee times with TE buffer. The beads were resuspended in 100 µl TE buffer containing 50 µg RNase A (Sigma) and the RNA was digested at 37° C for 30 min. Then the beads were washed once more with TE buffer and resuspended in 150 µl TE buffer containing 0.5% w/v SDS and 100 µg/ µl proteinase K and the crosslinking was reversed at 65° C overnight. Next day the DNA was extracted once with an equal volume of SAGE solution, precipitated with isopropanol in the presence of 5 µg glycogen and resuspended in 50 µl water.

2.5.3. PCR amplification

The primers used for the amplification of the distal enhancer of *Oct4* from the immunoprecipitated DNA were 5'-TGACAGAGTGGAGGAAACG-3' and 5'-ACACACAGCTACACATAGCA-3'. The reactions were carried out in 40 µl volume with 1 U Taq DNA polymerase (Roche), 0.25 mM dNTPs (ABgene), 0.2 µM of each primer, in the reaction buffer provided by the manufacturer (contains1.5 mM $MgCl_2$) on 4 µl of immunoprecipitated sample. The PCR program was 92° C 20 sec, 62° C 30 sec, 72° C 20 sec. The reaction was paused after 30, 35 and 40 cycles and 10 µl were removed from the reaction and resolved in 3% agarose gel in 1x TAE buffer, with 0.5 µg /ml ethidium bromide. The image analysis and signal quantification was performed with the Scion Image software.

2.6. MBD affinity purification (MAP)

The MBD affinity purification is used for the enrichment of methylated CpG islands from genomic DNA and has been modified from Cross *et al.* (1994).

2.6.1. MBD and CxxC recombinant proteins

The plasmids carrying the MBD and CxxC constructs were kind gifts from R. Illingworth and Dr H. Jorgensen. The MBD protein used for the selective targeting of methylated CpGs had been cloned into the pet30b plasmid between the Nde I and EcoR I restriction sites (pet30MBD76-167) downstream of a T7 promoter. The MBD protein was derived from the MBD domain of human MeCP2 and consisted of amino acids 76-167 of the full protein, tagged carboxy-terminally with six histidines. The Nde I restriction site (CATATG) served as the initiation codon.

The CxxC protein used as a control in the EMSA, contains the third CxxC domain of the mouse MBD1 with some flanking sequence (amino acids 352-448 of the full-length long form of MBD1) that has been shown to specificaly bind non-methylated CpGs (Jørgensen *et al.* 2004). It is tagged with six histidines at the

amino-end and cloned into the pet30b plasmid between the Hind III and EcoR I sites, downstream of a T7 promoter (pet30bCxxC).

2.6.2. Expression and purification of the MBD and CxxC recombinant proteins

BL21 DE3 pLysS chemically competent cells, prepared by K. Adie, were transformed with either the pet30MBD76-167 or pet30bCxxC vectors. This *E. coli* strain contains an inducible T7 RNA gene under the control of the lacUV5 promoter. Induction with IPTG allows production of T7 RNA polymerase which then directs the expression of the target gene located downstream of the T7 promoter in the expression vector. The BL21 DE3 pLysS also carries the plasmid pLysS that constitutively expresses T7 lysozyme. T7 lysozyme is a natural inhibitor or T7 RNA polymerase and serves to minimize the basal expression level of potentially toxic gene products before induction, allowing tight control of expression. 1:4 of the transformation reaction was plated on LB-agar containing 50 μg/ml kannamycin (pet30b selection) and 34 μg/ml chloraphenicol (BL21 DE3 pLysS selection). A single colony was picked and was expanded to a 200 ml liquid culture in LB with antibiotics at 37^{o} C over night with vigorous agitation. The overnight culture was diluted 1:50 in fresh LB broth (usually 6 lt) with antibiotics and was allowed to grow at 37^{o} C until it reached an OD_{600} 0.45-0.5. At this OD_{600} the bacteria are in their exponential phase of growth, in their optimum for protein production. When the culture reached the desired OD_{600}, IPTG was added to a final concentration of 1 mM and the cells were allowed to grow at 37^{o} C for another 2 hours. Previous optimization of the induction had shown that this period is optimal for good protein production.

The cells were spinned at 4200 rpm for 30 min at 4^{o} C and were subsequently washed twice with ice-cold PBS. Next, the cells were drained from excess buffer and frozen at -80^{o} C for 3 hours. The freeze-thawing is causing the fracture of the thick bacterial cell wall and aids lysis. The cells were thawed in ice and resuspended in 30 ml of Bacterial Lysis buffer per initial 1 lt of culture. 30 mg of lysozyme were added in 30 ml of lysate and the lysis of the cells was allowed for 30 min on ice. The cells where further disrupted with sonication with a Branson Sonifier S-450A, for 5 min at

30% output on setting 4, on ice. The lysate was precipitated at 4° C at 17,000g for 20 min and the supernatant was transferred to new tubes.

The cell lysate was next mixed with 1ml Ni-NTA sepharose beads (Amersham) for every initial 1.5 lt of culture and left at 4° C for 2 h with agitation. The Ni-NTA beads had been previously equilibrated in Bead Wash buffer. After this period the beads were sedimented at 4° C and resuspended in 5 bead volumes of Bead Wash buffer. The beads were washed a further twice and the protein was eluted with one bead volume of Elution buffer for 5-7 times. The efficiency of the purification was assed with SDS-PAGE gel electrophoresis. The protein-containing fractions were pooled and dialysed twice in 2 lt of Dialysis buffer over night. The dialysed protein was centrifuged to remove any non-soluble material. The protein concentration was calculated from the A_{260} measurement with a NanoDrop ND-1000 spectrophotometre and stored at 4° C for up to one week.

2.6.3. SDS-PAGE (polyacrylamide gel electrophoresis)

The 15% acrylamide gel contained 0.4 M Tris-HCl (pH8), 0.1% w/v SDS, 0.1% w/v APS and 1 µl/ml TEMED, added in this order. The 5% stacking gel contained 0.125 M Tris-HCl (pH6.8), 0.1% w/v SDS, 0.1% w/v APS and 1 µl/ml TEMED. Samples from the induction and purification steps were heated in 1x Laemmli buffer at 100° C for 10 min and then cooled in ice before loading them in the gel. The electrophoresis was performed in 1x Tris-Glycine at 200V until the bromophenol blue front reached the bottom of the gel. The resolved protein was visualised by immersing the gel for 30 min in Coomassie blue stain and destaining over night in Coomassie destaining solution.

2.6.4. Electrophoretic mobility shift assay (EMSA)

The methyl-CpG binding activity of the purified recombinant proteins was assessed with EMSA. No protein, 25 ng, 50 ng, 100 ng and 200 ng of MBD protein or 750 ng of control CxxC protein were incubated in room temperature for 5 min in 1x Binding buffer and 1 µg of the non-specific competitor dAdT (Amersham) at a final volume of 18 µl. After this short pre-incubation period, 2 µl of methylated or unmethylated radioactive CG11 probe were added to the mixture and the complexes were allowed to form for 25 min in room temperature. The CG11 probe (Meehan *et*

al. 1989) is 135 bp long and contains 27 CpGs (20 Hha I and 7 Hpa II sites). It was a kind gift from R. Illingworth, methylated with Hha I and Hpa II methylases and end-labelled with P^{32}.

The formed complexes were resolved with electrophoresis in 1.3% agarose in 0.5x TBE at 120 V for 4.5 h in 4^{o} C. The gel was dried under vacuum at 80^{o} C, exposed to a phosphor screen (Amersham) for 2 h and the signal was scanned with a STORM Phosphorimager (Amersham).

2.6.5. Packing of the column

40-60 mg of recombinant protein were bound to 1.2 ml of Ni-NTA sepharose beads (Amersham) for 2 h as described previously (section 2.6.2). The beads were then washed at 4^{o} C twice with two bead volumes of 1:9 Column buffer A:Column buffer B, twice with two bead volumes of 1:9 Column buffer A:Column buffer B as before but in the presence of 10 mM imidazole and again another twice with two bead volumes of 1:9 Column buffer A:Column buffer B. Finally the beads were resuspended in 10 bead volumes of 1:9 Column buffer A:Column buffer B and packed onto a 1 ml Tricorn 5/50 (GE Healthcare) FPLC column at a flow rate of 1ml/min using a peristaltic plumb. Each column was stored at 4^{o} C and was used for approximately ten runs.

2.6.6. Preparation of genomic DNA for MAP

The brain genomic DNA was extracted as described in section 2.2.2 and the ES (E14) and RA10 DNA was extracted as described in section 2.2.1. 100-130 µg of high molecular weight genomic DNA were digested with 250 U of Mse I (New England Biolabs) in a reaction volume of 200 µl at 37^{o} C for two hours in the buffer recommended by the manufacturer. To ensure complete digestion of the DNA, another 250 U of enzyme were added after two and four hours and the reaction was allowed to proceed over night. Mse I (TTAA) is a frequent cutter that preferentially recognizes sequences outside CpG islands. With this first step many AT-rich non-CpG island-like sequences are destroyed and the genome is divided into predictable Mse I fragments. After the end of the reaction, a fraction of the digested DNA was resolved in 1% agarose gel in 1x TAE buffer, with 0.5 µg /ml ethidium bromide to ensure complete digestion.

To prevent self-ligation later, the digested DNA was dephosphorylated with 80 U Calf Intestinal Phosphatase (New England Biolabs) at 37^{o} C over night, according to the manufacturer's instructions. To assess the efficiency of the phosphatase treatment, a small fraction of the dephosphorylated DNA was incubated with 1U T4 DNA ligase (New England Biolabs) for 1 hour and resolved in 1% agarose gel in 1x TAE buffer, with 0.5 μg /ml ethidium bromide. The protocol was continued only if there was no shift in the average fragment size in comparison to the non-ligated digested DNA.

Next, the DNA was ligated with the adaptors that would later be used for its amplification. The adaptors were ordered as complementary oligonucleotides: 5'-GGTCCATCCAACCGATCT-3' and 5'-Pi-TAAGATCGGTTGGATGGACC-3'. They were mixed in equimolar concentrations, boiled for 10 min and allowed to slowly cool down at room temperature. This enabled their annealing into the double-stranded adaptor, phosphorylated at the 5' of the "sticky" end:

5'-GGTCCATCCAACCGATCT-3'
3'-CCAGGTAGGTTGGCTAGAAT-Pi-5'

For, the ligation, the digested and dephosphorylated genomic DNA was mixed with an excess (10 μmol) of the adaptor and 3,200 U of T4 DNA ligase (New England Biolabs), in the buffer supplied by the manufacturer. The reaction was carried out in a final volume of 500 μl, at 16^{o} C, overnight.

At the end of the ligation reaction the solution was expected to contain mainly, Mse I-digested genomic DNA ligated to an adaptor at each end, free adaptors and self-ligated adaptors. The free adaptors and self-ligated 38-mers were removed with the QIAquick PCR purification kit (Qiagen). Since there were a big excess of adaptors in the reaction, Mse I-digested genomic DNA ligated to an adaptor at only one end and non-ligated Mse I-digested genomic DNA were not expected in significant amounts unless the ligation had been unsuccessful. To check the ligation efficiency, 20, 50 and 100 ng of ligated DNA were amplified with 0.4 μM universal primer (5'-GGTCCATCCAACCGATCTTA-3'), 2.5 mM $MgCl_2$, 0.4 mM dNTPs (ABgene) and 1.5 U Red Hot Taq polymerase (ABgene), in a final volume of 25 μl, in the buffer supplied by the manufacturer. The cycling program was: 95^{o} C 2min, 30x(95^{o} C 50 sec, 58^{o} C 50 sec,72^{o} C 3 min), 72^{o} C 7 min. The reactions were resolved in 1% agarose gel in 1x TAE buffer, with 0.5 μg /ml ethidium bromide and

were inspected for efficient and comparable amplification of all DNA dilutions. The successful ligation reactions were precipitated, a small fraction was kept as the input and the remaining was resuspended in 1:9 Column buffer A: Column buffer B before running it through the MBD column.

2.6.7. MBD affinity chromatography

The methyl-CpG island purification was performed on an ACTA purifier (GE Healthcare). The DNA sample was injected in the column and the methyl-CpG islands were eluted with a linear NaCl gradient (0.1 to 1 M, mix of Column buffers A and B) in 3ml fractions. A graphical overview of the elution program is given in Figure 2-1. To evaluate the degree of separation and to identify the methyl-CpG island containing fractions, 300 µl of each fraction were precipitated and resuspended in 60 µl of TE buffer. Equal amounts of each fraction were PCR

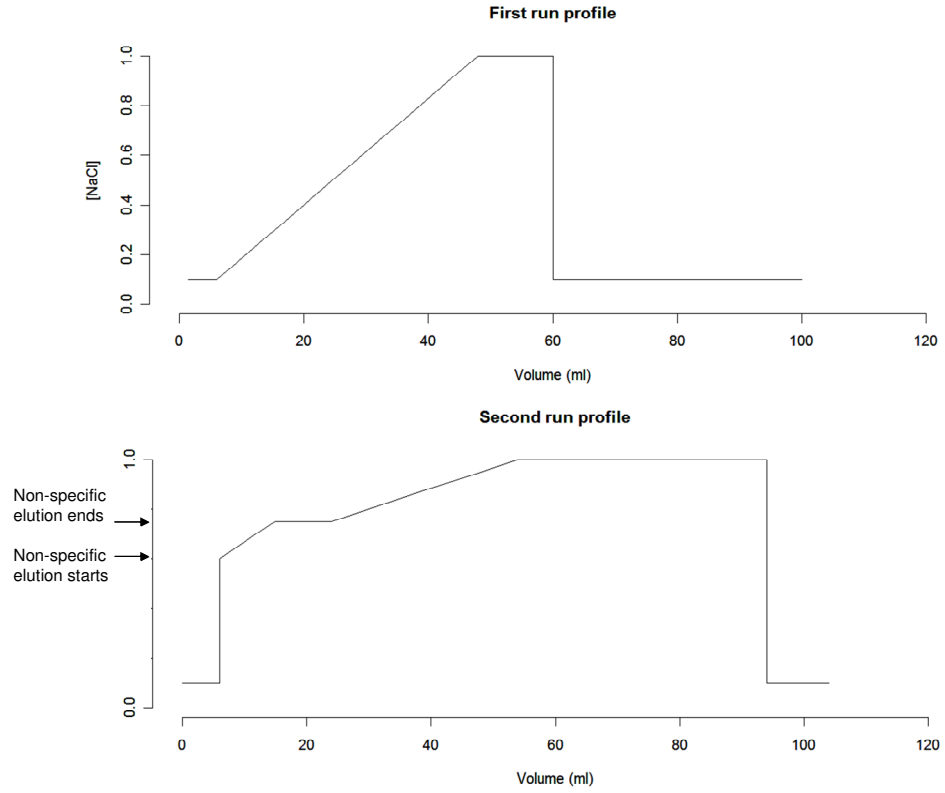

Figure 2-1. Column run profiles. Plots showing the salt gradients used for the first and second column runs. The first gradient was always the same while the second depended on the salt concentration that was required for the non-specific and specific elution of DNA.

amplified using the diagnostic primers listed in Table 2-5. All the primers were designed so that they are not interrupted by an Mse I recognition site and they were specific for differential methylated regions (DMRs) of known methylation status in males and females. 2 µl of each fraction were amplified with 0.3 µM of each primer, 2.5 mM $MgCl_2$, 0.4 mM dNTPs (ABgene), 5% v/v DMSO and 1.5 U Red Hot Taq polymerase (ABgene), in a final volume of 30 µl, in the buffer supplied by the manufacturer. The cycling program was: 95° C 3min, 30x(95° C 30 sec, 58° C 30 sec,72° C 30 sec), 72° C 10 min. The reactions were resolved in 2% agarose gel in 1x TAE buffer, with 0.5 µg /ml ethidium bromide.

The fractions that where identified to contain the methylated CpG islands were pooled, diluted to 0.1 M NaCl with Column buffer A and were passed through the column for a second time. The precise salt gradient of this second run was determined by the salt concentrations that appeared to elute DNA non-specifically and specifically in the first run (Figure 2-1). In more detail, it consisted of a linear NaCl gradient between the lowest and highest non-specific elution concentrations, then a wash at the highest salt concentration that does not elute methylated CpG islands and then linear gradient of the higher NaCl concentrations for the specific elution. Finally, the column was washed from remaining DNA with a long wash at 1 M NaCl and re-equilibrated at 0.1 M NaCl. As before, the fractions were tested by PCR with the same diagnostic primers to identify the ones that contain the specific eluate that were pooled, precipitated and resuspended in 300 µl TE buffer.

2.6.8. Preparation of the affinity-purified methyl-CpG islands for array hybidization

The MAP-purified methyl-CpG islands were amplified with the Universal primer described previously (section 2.6.6). In more detail 2 µl of the MAP-purified sample and 2 µl of 1:100 dilution of the input (pg range) were PCR amplified with 1.8 µM of the primer, 2.5 mM $MgCl_2$, 1 mM dNTPs (ABgene), 5% v/v DMSO and 1.5 U Red Hot Taq polymerase (ABgene), in a final volume of 50 µl, in the buffer supplied by the manufacturer. The annealing temperature in the cycling program was 58° C and the elongation time 4 min. The number of cycles was kept in the linear range of amplification. The amplified DNA was cleaned from non-incorporated nucleotides, primers, salt and protein with the QIAquick PCR purification kit

(Qiagen), adjusted to a concentration of 250-500 ng/µl and send to Nimblegen (NimbleGen Systems, Inc.).

Table 2-5. Diagnostic primers used in MAP

DMR-gene	Forward primer	Reverse primer
Igf2r	TATCGGCCCTCGTGTAGTTC	GAGGATTCCACGCGTTAGAG
HPRT	TCAGGCCCACCTAGTCAGAT	CGGAAAGCAGTGAGGTAAGC
Xist	AATTAGGACACCGAGGAGCA	TACGAGCACTCCTTGGCTTT
Ccne	GCTGGTCCACAGGAGACCTA	CACTGTCCCTCCTGACTCGT
Ddx 4	CTGGAGCGGAGAGGTGAGT	GCCTCAGGCCTTCACACC

2.6.9. Preparation of the custom-made mouse CpG island oligonucleotide tiling array

The database mining for the selection of the mouse genomic regions to be included in the array was carried out by Dr Alastair Kerr. Briefly, the mouse genome (NCBI build 34) was *in silico* digested with Mse I and repeat masked. Then, all the Mse I fragments that contained less than 100 nt of informative sequence or more that one third of low complexity sequence were removed. Of the remaining approximately three million Mse I fragments, only the ones that fulfilled our CpG island criteria, *i.e.* observed versus expected CpG ratio (o/e) equal or greater than 0.6 and GC content equal or greater than 50%, were kept. The CpG island criteria were applied in a window of 500 nt with a 50 nt step. Taking into consideration that the DNA had been fragmented, neighbouring Mse I fragments that fulfilled the criteria only when taken together were also included. The selected TTAA-flanked CpG island sequences were sent to the Nimblegen design team were they were scanned once more for regions of low complexity. The probes for each Mse I fragment were designed to be isothermic, tiled at 48 bp. At the end of this process the mouse CpG island array contained 26,687 TTAA-flanked sequences, represented by 385,215 probes.

2.6.10. Identification of the CpG islands

The CpG islands were assempled from the Mse I fragments by grouping together those fragments that were separated by less than or equal to 200 nucleotides

of genomic sequence. The Mse I fragments that were grouped in this way were further screened according to their length and if two or more of them were covering a sequence less than 2500 nucleotides long then they were all grouped together as one CpG island. After this process, the 26,687 Mse I fragments were assembled into 20,755 CpG islands. For the calculation of the position of a CpG island relative to a gene, those that were on the 5' untranslated region, the first inton or the first exon were termed 5'-associated. Those that were on the 3' untranslated region consisted the 3'-associated CpG islands and those that were not in the vicinity of any gene were intergenic. The remaining CpG islands that did not classify in any of the categories defined above, were characterised as intragenic.

2.6.11. Normalisation and pre-processing of the array data

The returned array data were analyzed using the package Limma (Linear models for microarray data) (Smyth 2004) of R in Bioconductor (http://www.bioconductor.org).

The arrays were normalised using a variation of the composite Loess normalisation method based on observed data as suggested in Wang *et al.* (2002). The invariant probes used for the calculation of the Loess regression line were identified by having an M value ($\log_2$ of the two channel ratio) between -0.5 and 0.5, *ie.* very close to zero, after global Loess normalisation and it was ensured that the selected probes spanned the entire intensity range. After determination of the invariant probes, the raw hybridisation data were normalised by fitting them to the Loess regression determined by the invariant spots. Then statistical analysis of the microarray data was performed with an empirical Bayes model. A moderated t-statistic test for differential enrichment was computed for each contrast for each gene. The confidence values produced by the t-statistic were expressed as q values for false discovery rate (Storey and Tibshirani 2003).

2.6.12. Real-time qPCR verification of the MAP results

All the primers used for the real-time quantitative PCR (qPCR) on genomic DNA that was purified with the MAP method are shown in Table 2-6.

Table 2-6. Primers used in qPCR for the verification of microarray data.

Region	Mse I fragment	Forward primer	Reverse primer
Br1-1	CHR1:6923991-6924628	GGCAGTGCATAGTGGGACT	ACCCTGGTGAAGAACTGCATTG
Br1-2	CHR1:186740536-186741471	GTCAGGGTAGGCTAAGGTGAG	TTCCACACAGGACTTCAGCATTT
Br2-1	CHR2:32266324-32267646	CTGTGGGTGACAGGATCT	CGACGAGACGCACACTCAAG
Br2-2	CHR2:74593463-74594377	CTGCTATTGTGACCTTCC	TCGAGGCTTGCACACAGTCA
Br2-3	CHR2:166928637-166929536	GAGAAGCCCAACTCATCGG	ACTCATCGTGAGGAGCCAGGAG
Br4-2	CHR4:88192272-88195185	GCCGTTCTCTCTCCTGTCA	TCCCTTCCTAAGGCCCTCTCTA
Br4-3	CHR4:134596644-134597844	GACTCTGGCTGGGGTAGAGAA	GCCCAAGCTCCACATTGTCA
Br5-1	CHR5:113129180-113130846	GAAATGTGTAGAGCAAGC	AGGCTTTCTCTCCGTCGTGGT
Br7-1	CHR7:4601268-4609003	CTTCACTTAGCACCATTGC	TATTGTATGTTTGTGCGCCAGTG
Br7-3	CHR7:21794900-21796927	GGACGACAAGGGCTATTTC	GCTGGACCAGATGTGAGGACTG
Br9-1	CHR9:121196775-121197768	GTTCAGAAAAGGGTCTGTTAGC	CTGAGCAGTAGACAGGCGGTGT
Br9-3	CHR9:121487252-121490211	CTCTCCCCACAGGCACTT	GCAGGAAGGCTGGTATCGTTC
Br10-1	CHR10:30622534-30623567	GGGAAGCGGGTGCTAGAG	AGTCCAGCTCCCTTCCAGTCG
Br11-1	CHR11:100683807-100684882	AAGATCATCCCTGGCTCT	AAGCACAAGAACCCTGATGTCC
Br15-1	CHR15:76258629-76266599	CAGCCAAAGTAGGAACAGC	CTCCCTCCTCTTTGCACCTGTC
Br15-2	CHR15:80455703-80457753	CTGGGTCTTGGTGCTAGATTAT	GGAGCTGGACTGCGGAAAGAC
Br16-1	CHR16:22107956-22108336	GGAAGCCGGACTTGATTT	ACGTGCCACAGTCGATAAGCAT
Br16-2	CHR16:93750542-93751116	TCCGATGTGACACACTTAC	CAACCTTCCCTCTCATGGACTTC
Br17-1	CHR17:33949358-33951691	GACCTCTAATCATTCGCTTTA	CTGGTTCCCTCCGAAGTTTCC
Br17-2	CHR17:5286533-5288507	TACCCGTGGAGCTTAGAGG	CGACACTACCGAGCACATCCAG
Br18-2	CHR18:37351238-37352485	AAGGGACTACGAACTCCACC	GGGCCGTAGATGAGGACTCTG
BrX1	CHRX:68614710-68616575	TAGTCAGGTGTGAGGTGTC	AAGCTGGCCGCTGGTATCCT
BrX2	CHRX:69632339-69632911	AGCCAGAAGAAAATGCAC	ATGCTCTCACGGTCCAGGAAT
BrX3	CHRX:12278734-12280449	GAGCACGCCGTCTAAGAA	GTCGCCTCTACTCCCAAGCTG
Dev1-2	CHR1:6383009-6385694	TCTGGCAGCTCTTGAAGA	GGACTCCATGTCAGTCTCAACTCC
Dev6-1	CHR6:17780608-17781540	TGGCTCCGTAAGTCGTCT	CCTCCGCTGAGATACTGACATCC
Dev7-2	CHR7:13920979-13921968	CATCCTTTGTGAGGTAAACC	GGGAGGGACACGAGAAGACAGA
Dev8-1	CHR8:118186994-118187994	GGAGGAGTTAGCGAGCCA	TAGGTGTCGGCTATGTGCTTGG
Dev10-1	CHR10:120756802-120757699	ACAAAAGAAACAGGAAGGG	ACAAGTCCGGCATTGGGAAC
Dev10-2	CHR10:14224996-14225901	TCAGCAGGTGCCCAGTC	TTGCTCAACAAATATAAGGCATGA
Dev10-5	CHR10:59389229-59390186	CCCAATCTTGTTGGCAACGTC	AGTCAATACCCTGATAAGGCC
Dev11-2	CHR11:106127743-106128527	CACAGCCAATGAGCATGA	CTTCACCTTCAGCAGCCTCACT
Dev12-1	CHR12:52087446-52087926	CTTGCGCTCCAGAGTCTG	ACCACAGCCGAGGGACTCTT
Dev14-2	CHR14:40176456-40177819	CTCTATGACATGCACAGG	AGGGACCCGTGCCTGTAGTCT
Dev17-1	CHR17:85056684-85057302	TTGAACAAACTTGGGTGC	AGAGGGCAAAGCGCAACTCT

24 Mse I fragments that showed various distributions of their M values in the brain microarray experiment were selected and qPCR primers were designed for them. qPCR was performed on 10 ng of the MAP-enriched and input PCR-amplified DNA from two female and one male samples (six female and three male individuals). The PCR reactions were performed in triplicates under similar conditions as described in section 2.4.4. The formula used for the calculation of the enrichment of each Mse I fragment was:

$$2^{-(\dot{C}_{T(MBD)}-\dot{C}_{T(IN)})}$$

where $\dot{C}_{T(MBD)}$ is the mean threshold cycle of the MAP-enriched samples and $\dot{C}_{T(MBD)}$ is the mean threshold cycle of the input samples.

The Mse I fragments that were consistently enriched in the RA10 samples but not in the samples from ES cells were identified according to the following criteria: M value equal to or less than 0 in the ES microarrays and equal to or less than 2 in the RA10 microarrays for more than 90% of the probes of each Mse I fragment. This process identified 24 potentially enriched Mse I fragments. For the verification of the enrichment, 12 fragments were randomly selected from the 24 identified by the microarray experiment and qPCR primers were designed for them. qPCR was performed on equal volumes of the ES and RA10 DNA. Equal quantities of input DNA from these two sources had been used for MAP purification, so PCR on equal volumes of the MAP-purified DNA could determine whether a fragment was enriched in one sample relative to the other. The PCR reactions were performed in triplicates under similar conditions as described in section 2.4.4. The formula used for the calculation of the enrichment of each Mse I fragment was:

$$2^{-(\dot{C}_{T(RA10)}-\dot{C}_{T(ES)})}$$

where $\dot{C}_{T(RA10)}$ is the mean threshold cycle of the RA10 samples and $\dot{C}_{T(ES)}$ is the mean threshold cycle of the ES cell samples.

2.7. Solutions

The solutions used in all the protocols described are listed here in alphabetical order.

Bacterial Lysis buffer: 50 mM sodium phosphate buffer (pH 8), 300 mM NaCl, 10% v/v glycerol, 10 mM Imidazole, 15 mM β-mercaptoethanol, 0.5 mM PMFS

Bead Wash buffer: 50 mM sodium phosphate buffer (ph 8), 300 mM NaCl, 10% v/v glycerol, 20 mM Imidazole, 15 mM β-mercaptoethanol, 0.5 mM PMFS

1x Binding buffer: 6 mM Tris-HCl (pH 8), 6 mM $MgCl_2$, 150 mM NaCl, 3% glycerol, 1 mM DTT, 10 ng/µl BSA

Bisulfite solution: 3.1 M sodium bisulfite dissolved in the presence of 0.6 M NaOH, 0.5 mM hydroquinone previously dissolved in water at 55^o C, pH 5

Buffer III: 0.25 M LiCl, 1% v/v NP-40, 1% w/v deoxycholate, 1 mM EDTA, 10 mM Tris-Hcl (pH 8.1)

Church and Gilbert buffer (modified): 0.5 phosphate buffer (pH 7.2), 10 mM EDTA, 7% w/v SDS

Column buffer A: 20 mM Hepes (pH 7.9), 0.1 % v/v Triton X-100, 10% glycerol, 0.5 mM PMSF, 10 mM β-mercaptoethanol

Column buffer B: 1M NaCl, 20 mM Hepes (pH 7.9), 0.1 % v/v Triton X-100, 10% glycerol, 0.5 mM PMSF, 10 mM β-mercaptoethanol

Coomassie blue stain: 45% v/v methanol, 10% v/v acetic acid, 1g/lt coomassie blue

Coomassie destaining solution: 5% v/v methanol, 7.5% v/v acetic acid

Dialysis buffer: 50 mM sodium phosphate buffer (pH 8), 300 mM NaCl, 10% v/v glycerol, 15 mM β-mercaptoethanol, 0.5 mM PMFS

Elution buffer: 50 mM sodium phosphate buffer (pH 8), 300 mM NaCl, 10% v/v glycerol, 250 mM Imidazole, 15 mM β-mercaptoethanol, 0.5 mM PMFS

1x Laemmli buffer: 62.5 mM Tris-HCl (pH 6.8), 10% v/v glycerol, 1.25% w/v SDS, 2.5 % v/v β-mercaptoethanol, bromophenol blue

Luria-Bertani (LB) broth: 10 g/lt bacto-tryptone, 5g/lt bacto-yeast extract, 10g/lt NaCl

LB-agar: LB broth with 1.5% agar-agar

Lysis buffer I: 0.5 M Tris-HCl (pH8), 20 mM EDTA, 10 mM NaCl, 1% w/v SDS, 0.5 mg/ml proteinase K

Lysis buffer II: 50 mM Tris-HCl (pH 7.5), 5 mM EDTA, 100 mM NaCl

1x MOPS buffer: 20 mM 3-[N-morpholino]propanesulfonic acid, 5 mM sodium acetate, 1 mM EDTA, pH 7

NE1 buffer: 20 mM HEPES (pH 7), 10 mM KCl, 1 mM $MgCl_2$, 0.5 mM DTT, 0.1% v/v Triton X-100, 20% v/v glycerol, 200 mM NaCl, protease inhibitor cocktail (Roche)

PBS: 137 mM NaCl, 10 mM Phosphate (pH 7.4), 2.7 mM KCl

RA stock solution: 100 mM all-trans retinoic acid in DMSO

1x RNA loading buffer: bromophenol blue, 30% v/v formamide, 20% v/v glycerol, 4 mM EDTA, 0.88 M formaldehyde, 0.4x MOPS buffer

SAGE solution: 25:24:1 v/v/v phenol:chlorophorm:isoamyl alcohol (pH 8)

10x SSC buffer: 0.15 M trisodium citrate (pH 7), 1.5 M NaCl

1x TAE buffer: 0.04M Tris-acetate (pH 8.3), 2mM Na_2EDTA

0.5 TBE buffer: 445 mM Tris-HCl, 1 mM Na_4EDTA, 445 mM boric acid

TE buffer: 10mM tris (pH 7.5), 1mM EDTA

1x Tris-Glycine buffer, 3g/lt Tris-HCl, 14.4g/lt glycine, pH 6

TSE I: 0.1% w/v SDS, 1% v/v Triton X-100, 2 mM EDTA, 20 mM Tris-HCl (pH 8.1), 150 mM NaCl

TSE II: 0.1% w/v SDS, 1% v/v Triton X-100, 2 mM EDTA, 20 mM Tris-HCl (pH 8.1), 500 mM NaCl

3. The role of DNA methylation in early development through regulation of the pluripotency transcription factor OCT4

3.1. Pluripotency transcription factors

Transcription factors that are necessary for the maintenance of pluripotency are termed pluripotency transcription factors. These factors are downregulated as the cells differentiate and thus are not the general transcription factors or co-factors that are expressed in many cell types. They are experimentally identified as the products of genes that, when knocked down, prevent maintenance of the stem cell character of the cells in culture. In their absence, expression of stem cell markers such as alkaline phosphatase is lost, differentiation markers are induced and/or the typical embryonic stem (ES) cell morphology changes to more flattened, larger cells that grow as a monolayer.

The first pluripotency factor identified was OCT4 (Niwa *et al.* 2000; Nichols *et al.* 1998), the product of the *Pou5f1* gene, while now NANOG (Chambers *et al.* 2003) and SOX2 are also widely accepted as important pluripotency factors. The list is slowly growing with ESRRB, RIF1, FOXD3 and TBX3 (Ivanova *et al.* 2006; Loh *et al.* 2006; Hanna *et al.* 2002) but more studies are required to establish the role of these factors in self-renewal and pluripotency. Pluripotency transcription factors are very interesting as they suggest the possibility of a rather simple way to de-differentiate somatic cells just by forcing their re-expression (Okita *et al.* 2007; Takahashi and Yamanaka 2006; Wernig *et al.* 2007). These reverted pluripotent cells could then be moulded into any cell type needed in research and medicine. Identification of the targets of these factors provides further understanding of how pluripotency and differentiation is brought about, while elucidation of how these transcription factors are themselves regulated can answer fundamental questions about the timing and specificity of gene transcription.

3.1.1. Targets of pluripotency transcription factors

Chromatin immunoprecipitation of OCT4, NANOG and SOX2 coupled with microarray hybridisation (Boyer *et al.* 2005) or DNA-tag sequencing strategies (Loh *et al.* 2006) has shown that they share a large proportion of their target genes. These target genes in mouse can be classified as genes involved in transcription, morphogenesis, organogenesis, development, and metabolism. In human, a

prominent class of genes regulated by these transcription factors is the homeodomain gene family (Boyer *et al.* 2005). Only 9.1% of OCT4-bound genes and 13% of NANOG-bound genes are common to both human and mouse. This could be attributed merely to the different experimental approaches, or could imply that there are different mechanisms for conferring pluripotency in mouse and human. An interesting outcome of the analyses presented here is that pluripotency transcription factors seem to invariably regulate both themselves and each other, forming a complex regulatory network of feedback and autoregulatory loops.

OCT4 and NANOG appear to bind discrete regions in the shared promoters (Loh *et al.* 2006), while OCT4 and SOX2 form a heterodimer and synergistically bind to their targets (Nishimoto *et al.* 1999; Yuan *et al.* 1995). Correspondingly, NANOG appears to be able to form a heterodimer with another protein, SALL4, and co-occupy at least some of the NANOG target genes (Wu *et al.* 2006).

3.1.2. Regulatory interactions of pluripotency transcription factors

As mentioned in section 3.1.1, a characteristic of the pluripotency transcription factors is that they regulate the transcription of themselves and each other. Chromatin immunoprecipitation with OCT4, NANOG and SOX2 antibodies revealed they interact with their own regulatory regions as well as each others' (Boyer *et al.* 2005). Depletion of *Oct4* with RNAi resulted in the downregulation of *Nanog* and similar depletion of *Nanog* resulted in the downregulation of *Oct4* (Loh *et al.* 2006). *Sox2*, *Esrrb* and *Rif1* were also downregulated in both cases. Moreover, there is clear evidence that the OCT4-SOX2 heterodimer is responsible for the activation of *Oct4* and *Nanog* (Okumura-Nakanishi *et al.* 2005; Rodda *et al.* 2005), and that OCT4 binding in the *Sox2* enhancer regulates its activity (Catena *et al.* 2004). Finally, chromatin immunoprecipitation and reporter assays have shown that NANOG is responsible for the activation of its partner, *SALL4*, and *vice versa*, probably as a NANOG-SALL4 heterodimer (Wu *et al.* 2006). A simple diagram of the regulatory interactions described here is depicted in Figure 3-1.

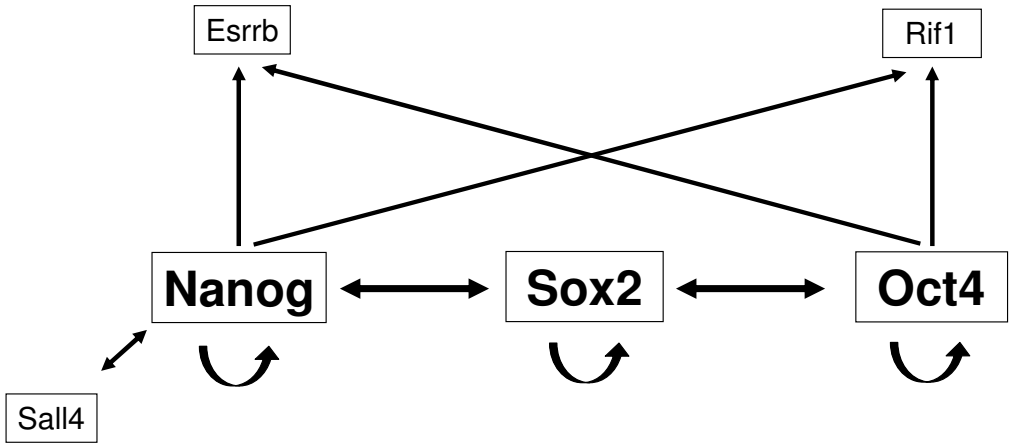

Figure 3-1. Known regulatory interactions of pluripotency transcription factors. The arrows symbolize transcription activation. Block arrows and letters represent well-studied interactions, whereas thin arrows and letters, interactions supported by limited evidence. The formation of heterodimers between pluripotency factors is not taken into account for this model.(Catena *et al.* 2004; Rodda *et al.* 2005; Okumura-Nakanishi *et al.* 2005; Loh *et al.* 2006; Wu *et al.* 2006).

3.1.3. The pluripotency transcription factor OCT4

OCT4 was first discovered by Schöler *et al.* (1989a; 1989b) as an octamer motif-binding activity present in progenitor germ cells, oocytes, ES cells and embryonic carcinoma (EC) cells . This binding activity was lost upon *in vitro* differentiation of ES and EC cells. Reporters under the control of an octamer motif-containing promoter were active in ES cells and became silenced as these cells differentiated . This reporter was specifically expressed in the inner cell mass (ICM) of the mouse blastocyst. These experiments were establishing what is now widely known: OCT4 is an octamer motif-binding transcription factor that is present in the oocyte, ICM and germ cells and its expression is lost as the cells differentiate.

The *Oct4* gene (ENSMUSG00000024406, Figure 3-2) is approximately 5Kb long, has 5 exons and encodes a protein of 376 amino acids. The first thorough deletion analysis for the characterisation of the *Oct4* regulatory region was conducted

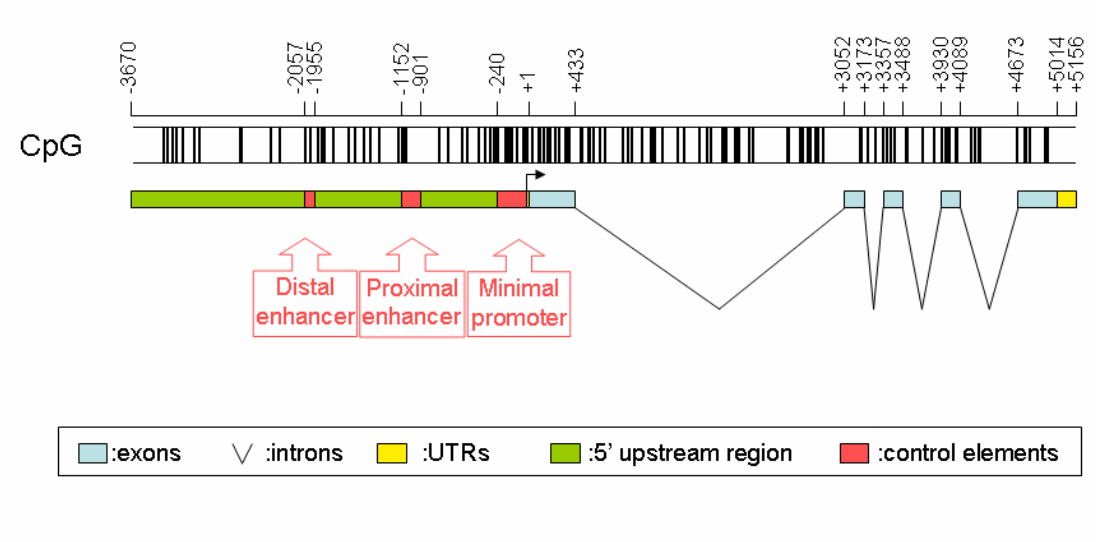

Figure 3-2. Structure of the *Oct4* gene and its upstream region. An alternative start site has been reported 63 bp downstream of the indicated start site. The longer transcript version is used for position numbering with reference to the transcription start site. The position of the CpG dinucleotides is shown. UTR: untranslated region. The map has been based in deletion analyses (Yeom *et al.* 1996) and the conservation of the regions between mouse, human and cow (Nordhoff *et al.* 2001).

by Yeom *et al.* (1996) and it identified a promoter (P), a proximal (PE) and a distal enhancer (DE). By introducing deletion constructs in mouse embryos, the authors showed that the promoter region is sufficient for reporter expression, the DE is necessary for expression in the ICM (and ES cells) and the germ cell lineage, while the PE is important for activation of the reporter in the post-implantation epiblast (as well as in embryonal carcinoma, EC, cells). This discrete role of the two enhancers suggests there must be a molecular "*switch from distal to proximal enhancer activity around implantation*" (Ovitt and Schoeler 1998).

In more detail, the promoter, which is approximately 60bp from the transcription start site, contains a putative Sp1/3 recognition site and a retinoic acid response element (RARE)-like sequence that partially overlap. It has been shown that the Sp1/3 element is occupied in undifferentiated ES and EC cells, but the occupancy is lost upon *in vitro* differentiation (Minucci *et al.* 1996). Furthermore, the

same study showed that an intact Sp1/3 site is necessary for the transcriptional activation of a reporter with the *Oct4* promoter. It is interesting that *Nanog* has also been reported to contain functional Sp1/3 binding sites in its promoter (Wu and Yao 2006). Nevertheless, in another study (Schoorlemmer *et al.* 1994), deletion of Sp1 had no effect in *Oct4* expression indicating functional overlap with Sp3 or some other GC-box binding transcription factor.

The GC-box binding orphan nuclear receptor SF1 has also been implicated in the activation of *Oct4* in EC cells through binding to the promoter at the RARE element (Barnea and Bergman 2000) and Fuhrmann *et al.* (2001) showed that displacement of SF1 by GCNF is necessary for *Oct4* repression. The interaction of GCNF with the *Oct4* promoter *in vivo* was confirmed by immunoprecipitation of the *Oct4* promoter with an anti-GCNF antibody shortly after RA-induced differentiation of ES and EC cells (Gu *et al.* 2005b). The same study showed that GCNF-/- ES cells failed to downregulate *Oct4, Nanog, Sox2* and other pluripotency markers upon induction of differentiation with RA, suggesting a more general role of GCNF in the regulation of pluripotency transcription factors. Except from the factors discussed here, many hormone receptors have been shown to bind the *Oct4* RARE *in vitro* and to exert some influence in the expression of reporter constructs containing the RARE element. These results however, are contradictory to the results of other researchers and there is no *in vivo* evidence for their support.

Much less is known about the factors that bind and regulate the activity of the two enhancers, both of which contain conserved RARE-like elements (Nordhoff *et al.* 2001). Chromatin immunoprecipitation with anti-LRH-1 antibody can precipitate the PE in ES cells (Gu *et al.* 2005a). Moreover, in concordance with the observation that the PE is important for maintaining *Oct4* expression at the epiblast stage, the same authors show that disruption of LRH-1 has no effect on *Oct4* expression at the inner cell mass (ICM), but causes dramatic reduction of its expression after implantation. They also show that there may be some functional overlap with the closely related receptor SF-1 and suggest a role of LRH-1 on promoter activation.

NANOG binds the entire region that spans the two enhancers, while OCT4 binds to the DE (Loh *et al.* 2006) but nothing is yet known about the functional significance of these interactions.

Interestingly, downregulation of DICER with siRNAs reduced the mRNA levels of *Oct4,* as well as *Nanog* and *Sox2* by about a half*,* giving rise to the argument that transcription of these pluripotency factors may also be regulated by RNA interference (Cui *et al.* 2007).

3.1.4. Epigenetic regulation of Oct4

Five years ago it was discovered that at least four CpG dinucleotides in the 5' upstream region of Oct4 quickly gain methylation after implantation (Gidekel and Bergman 2002). The same sites were unmethylated in EC cells and methylated in the trophoblast, pointing to some epigenetic regulation of *Oct4* expression. It was later shown that indeed *Oct4* can be reactivated in trophoblast and fibroblast cells after treatment with 5-aza-2'-deoxycytidine and trichostatin A (Hattori *et al.* 2004). Thorough methylation analysis showed that the *Oct4* promoter region is unmethylated in the cells where the gene is expressed and methylated where it is not. Furthermore, the active promoter is associated with H3 acetylation and this modification is lost in the silenced gene. Interestingly, H4 acetylation levels do not change.

The question was obvious, is methylation regulating *Oct4* expression or is it a secondary modification? Sato *et al.* (2006) investigated how the *Oct4* mRNA levels compare with the methylation levels in selected regions of P, PE and DE. They found that *Oct4* levels seemed to be well into decline before methylation appeared. Further investigation showed that the *Oct4* repressor GCNF co-immunoprecipitates with DNMT3a and DNMT3b and that GCNF and DNMT3a overexpression cause *Oct4* promoter hypermethylation. Detailed analysis of the methylation levels at the promoter (Gu *et al.* 2006) showed that embryos without a functional GCNF can not methylate the Sp1/3- RARE region of the promoter, although the methylation of the rest of the promoter is comparable to the wild-type, and causes the ectopic expression of the gene. They also showed that GCNF sequentially recruits MBD3 and MBD2 (in this order) to suppress transcription. As expected from the specificities of MBD2 and MBD3, GCNF and MBD3 binding is indifferent to methylation, whereas MBD2 requires methylation for binding to the *Oct4*. However, the need for DNA methyltransferases for the downregulation of *Oct4* remains controversial; Gu *et al.* showed that DNMT3a/b -/- ES cells fail to methylate the

promoter of *Oct4* upon differentiation although the gene can be downregulated. On the other hand, Jackson *et al.* (2004) observed the opposite: DNMT3a/b -/- cells can not downregulate *Oct4*. The main difference between these two observations was the strategy followed for the differenciation of ES cells and the downregulation of Oct4. While Jackson *et al.* (2004) differentiated the cells by simple LIF removal, Gu *et al.* (2006) differentiated the ES cells by LIF removal and addition of retinoic acid (RA). Another difference between the two approaches is that Jackson *et al.* (2004) emphasise that the mutant cell lina was of advanced passage number, something that they had previously shown to affect the global DNA methylation levels in these cells. Gu *et al.* (2006) on the other hand do not mention the passage number of the cells they used. An interesting observation made by Feldman *et al.* (2006) that the *Oct4* promoter in DNMT1-/- fibroblasts appears to be less methylated than in wild-type, indicating that DNMT1 could also be an important player.

Regarding the histone modifications associated with the active and idle *Oct4* promoter, Sato *et al.* (2006) showed that silencing is accompanied by a reduction in the levels of the active chromatin modifications acetylH3K9/14 and trimethylH3K4 but no significant alteration in trimethylH3K9 or trimethylH3K27. They observed a reduction at dimethylH3K9. However, Feldman *et al.* (2006) disagreed as they saw an increase in di- and tri-methylH3K9 levels at the promoter and it was initiated before the onset of transcriptional repression. They investigated this further and found that the H3K9 methyltransferase G9a is responsible for this modification. They hypothesised that methylation of H3K9 is needed for DNMT3 binding to the promoter and long-term silencing of *Oct4* but not for repression initiation. The link between the G9a-driven and the GCNF-induced silencing of *Oct4* still remains to be discovered.

3.1.5. Aims

Despite the recent intensive research on *Oct4* regulation there are still many questions that remain unanswered. The aim of this study was to investigate how and why DNA methylation is established at the *Oct4* upstream region. In more detail, the exact time when DNA methylation first appears with regards to the transcriptional status of the gene, as well as the pattern with which it is established was investigated. Appreciating that most of the previous studies on *Oct4* methylation were either of

low resolution or focused on the promoter of the gene, the entire upstream region with all three regulatory elements was examined on each CpG position during the course of the gene's downregulation. It was expected that, given the previous knowledge of the important upstream regions that are occupied at different differentiation stages, as well as the transcription factors that bind to the different regulatory elements, such a study would make clearer what the molecular trigger that leads to DNA methylation at this locus is. Finally, in order to achieve this goal, an appropriate differentiation model system is established for the investigation of *Oct4* downregulation.

3.2. Establishment of the *in vitro* differentiation system

In vitro differentiation of ES cells into embryoid bodies (EB) is a technique that models the events of early development and is the first step in many protocols for directed differentiation of ES cells into specific cell types. Leahy *et al.* (1999) have shown that all the main cell lineages are present in EBs and the timing of their appearance matches closely that of the early embryo. During EB formation, the cells initially aggregate into compact masses. Later, through a program of apoptosis and cell contact-mediated interactions very similar to those that take place during *in vivo* differentiation, they form a cavity that bears similarities to egg cylinder-stage embryos (Choi *et al.* 2005).

In vitro differentiation of ES cells has the obvious advantage that the time, cost and ethical issues associated with harvesting embryos are overcome. Moreover, it allows the easy production of large quantities of differentiated material, which furthermore can be of much purer cell lineage than in the embryo. On the other hand, in this system, the environmental differentiation stimuli that are present *in vivo* are missing; there is no maternal environment or embryonic trophectoderm, which in turn means that not all the events of early development can be recapitulated *in vitro*. Finally, it should not be forgotten that the starting point of differentiation is not the

inner cell mass (ICM) but ES cells, which, although are derived from it, are not equivalent to it. Dispite its drawbacks, *in vitro* differentiation of ES cells has been judged to be appropriate for the investigation of the mechanism of *Oct4* downregulation. *Oct4* is active in ES cells as it is in the ICM and silenced in EBs like in the differentiated embryo. The lack of maternal signals or trophectoderm does not appear to influence this process. Assuming realistically that there is only one mechanism for *Oct4* silencing, *in vitro* differentiation of ES cells provides a good system for the purposes of this research work.

There are two common methods to induce differentiation of ES cells. The simplest way is to remove LIF and allow the cells to grow in suspension. This method relies on blocking the LIF-STAT3 pathway that promotes pluripotency. The alternative method requires the addition of RA too. RA is an activator of morphogenesis, development, and cell differentiation through the activation of retinoid receptors. Both methods have been applied in the past for varied differentiation periods in the investigation of the *Oct4* regulatory mechanism, making it difficult to combine the information from different laboratories. Here the two methods are interrogated in terms of their potential for silencing and methylation of *Oct4*.

3.2.1. RA-induced differentiation of ES cells is the most efficient method for Oct4 silencing

E14 ES cells were differentiated by removal of LIF for three, seven, fourteen and twenty one days (EB3, EB7, EB14 and EB21). Alternatively, RA was added to the medium after three days of LIF removal and the differentiation continued for another two and four days (RA2 and RA4). Northern hybridisation of RNA extracted from each differentiation stage (Figure 3-3) showed that *Oct4* down-regulation was quicker and more efficient after RA treatment. Three days after LIF removal the *Oct4* levels had dropped by 20% and then stayed almost stable even after twenty one days of differentiation (samples EB3-EB21). At this stage the embryoid bodies were not looking healthy and the cells were dying. In the RA samples by contrast, the *Oct4* levels went down to 50% after only two days after addition of RA.

Next, the degree of DNA methylation in the *Oct4* promoter at each stage of differentiation in these samples was analysed (Figure 3-4). DNA extracted from tail

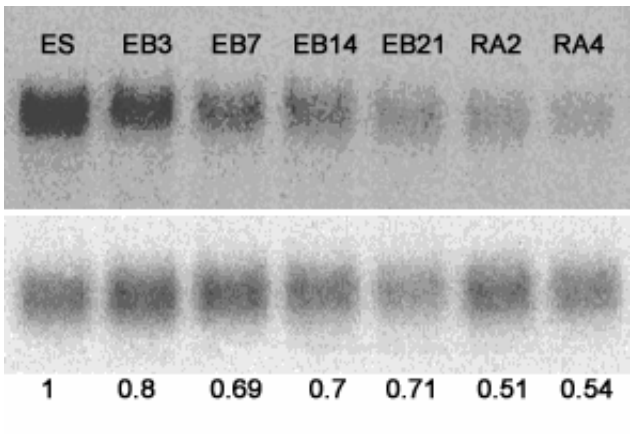

Figure 3-3. Expression analysis of *Oct4* during differentiation by LIF removal or addition of RA. E14 ES cells (ES) were allowed to differentiate *in vitro* by LIF removal for 3, 7, 14 and 21 days (EB3, EB7, EB14 and EB21), or, after 3 days of LIF removal, RA was added to the medium to a final concentration of 1μM for 2 and 4 days (RA2 and RA4). The top panel shows hybridisation against a probe for the entire last exon of *Oct4* and the bottom panel shows the control hybridization against S26. The numbers below the picture indicate the *Oct4* signal at each stage as a fraction of the signal in ES cells, after correction for equal loading.

tips was also analysed as a control of the natural methylation levels of DNA methylation at the *Oct4* promoter in finally differentiated cells. As expected, ES cells do not have any methylation, while methylation at the terminally differentiated tail cells is very high, reaching 45% (Figure 3-4, A). Differentiation by removal of LIF does induce some methylation establishment at the promoter but the process appears to be again very slow and inefficient and the methylation after twenty one days does not exceed 13% (Figure 3-4, B). On the other hand, addition of RA to the differentiation medium accelerates the methylation process and makes it more efficient; only four days after addition of RA (seven days without LIF) the cellsachieve 29% methylation (Figure 3-4, C). Interestingly, methylation is not gained gradually after RA addition. The promoter remains unmethylated for the first two days and then, within the next two days (approximately two cell duplications), methylation appears.

Because of these results, all the subsequent *in vitro* differentiation experiments were conducted with the RA method.

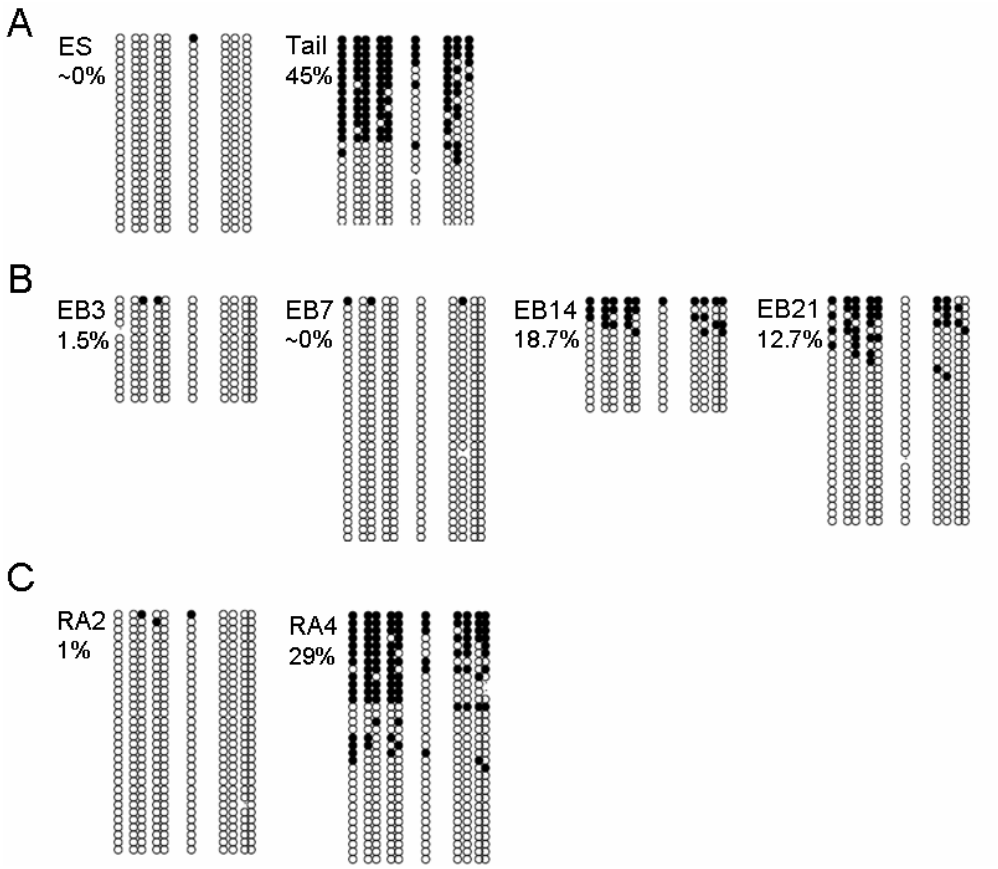

Figure 3-4. Methylation analysis of the promoter region of *Oct4* during differentiation by LIF removal or addition of RA. The region from -200 to +100 (+55) (left to right for each cell type) with respect to the transcriptional start site is analysed. Methylated and non-methylated CpGs are denoted by filled and empty circles respectively. The calculated percentage methylation is shown next to the data. (A) Methylation analysis of the *Oct4* promoter in pluripotent ES cells and terminally differentiated cells from the tail tip. (B) Methylation analysis during the course of *in vitro* differentiation of ES cells to embryoid bodies by LIF removal. (C) Methylation analysis during the course *in vitro* differentiation of E14 ES cells to embryoid bodies by addition of RA (after three days of LIF removal).

3.2.2. The RA-induced in vitro differentiation is reproducibly recreating events of early development

Before proceeding with examining the epigenetic regulation of *Oct4*, the reproducibility of the chosen differentiation program needed to be verified. Two sets of RA-induced embryoid bodies were produced from E14 ES cells (I and II) and examined with RT-PCR for the presence of various developmental markers (Figure 3-5).

Figure
3-5

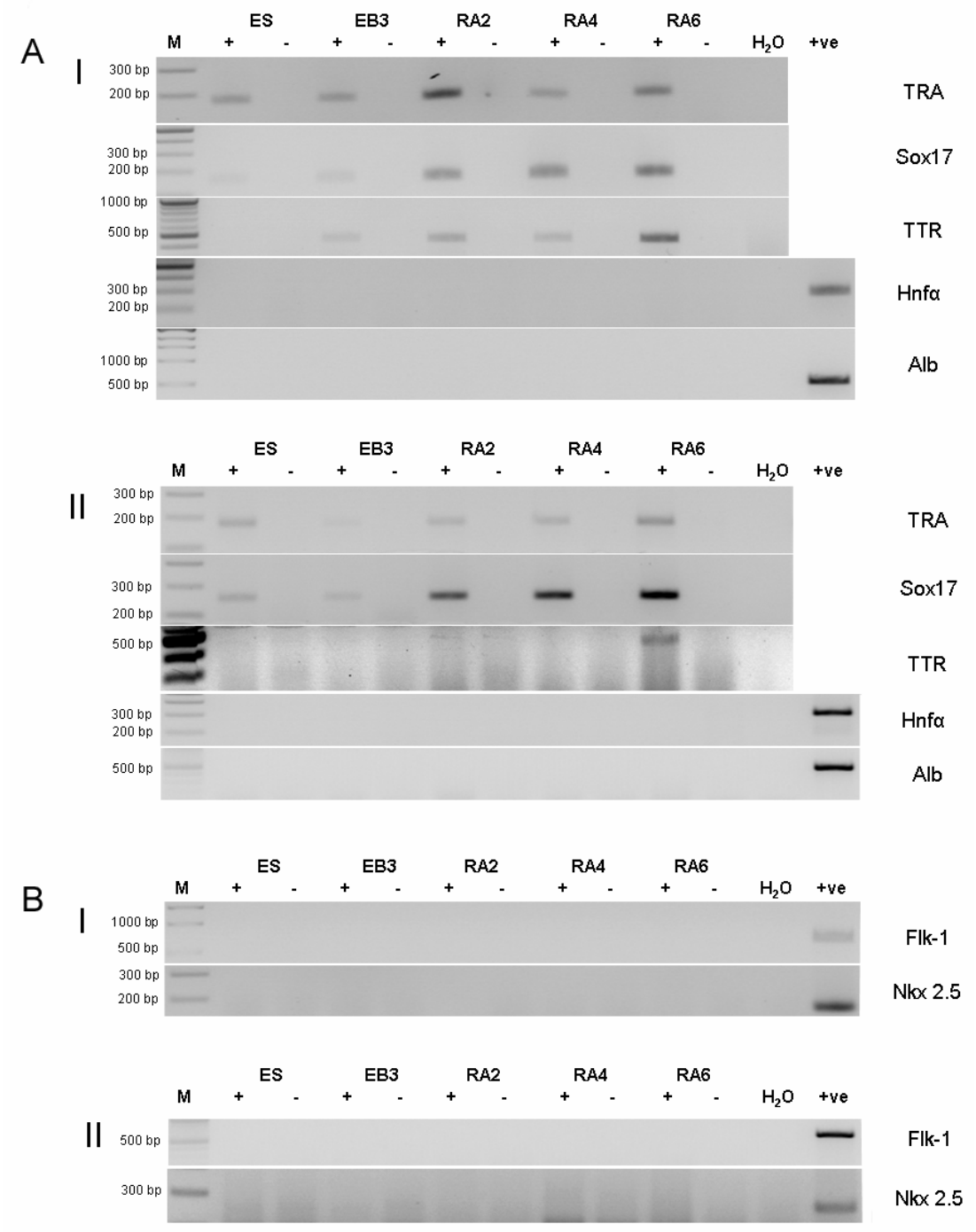

Figure 3-5. Expression of developmental markers during the course of *in vitro* differentiation of E14 ES cells. Two sets of embryoid bodies (I and II) were produced from E14 ES cells by removal of LIF for three days (EB3) and then addition of RA to the medium for the indicated number of days (RA2 to RA6). + : reverse transcribed RNA, – : mock-reverse transcribed RNA. Whenever a marker could not be detected in the embryoid bodies,

a positive control (+ve) was included to verify the reaction worked. A water negative control was included in all reactions. (Continues in next page)

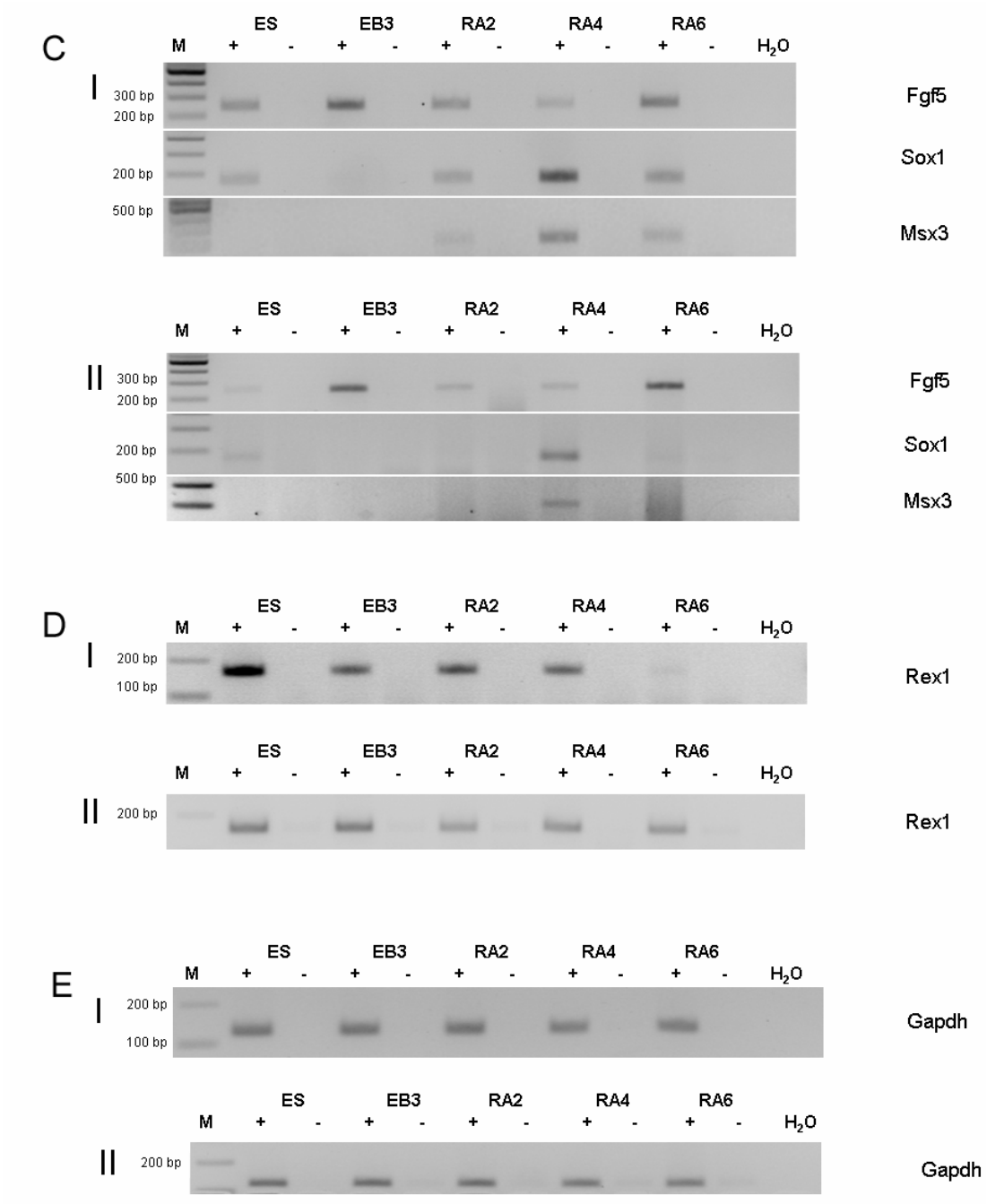

Figure 3-5 (Continued) M is the molecular weight marker. (A) Endodermal markers: transthyretin (TRA), SRY-box containing gene 17 (Sox17), transferring (TTR), hepatic nuclear factor 4 alpha (Hnfα), albumin (Alb). (B) Mesodermal markers: foetal liver kinase 1 (Flk-1), homeobox protein NK-2 homolog E (Nkx2.5). (C) Ectodermal markers: fibroblast growth factor 5 (Fgf5), SRY-box containing gene 1 (Sox1), homeo box msh-like 3 (Msx3). (D) Stem cell marker: RNA exonuclease 1 homolog (Rex1). (E) Loading control: glycerinaldehyde-3-phosphate dehydrogenase (Gapdh).

In both sets a variety of endodermal markers (Figure 3-5, A) was induced but not the hepatic marker *Hnfα* or the visceral endoderm marker *albumin*. None of the mesodermal markers (Figure 3-5, B) tested could be detected, while the primitive ectodermal marker (Figure 3-5, C) *Fgf 5* could be detected since early. Other ectodermal markers that are typical of the neurectoderm (*Sox1* and *Msx3*) could also be detected, but at later stages of differentiation. Finally, the ES cell marker *Rex1* (Figure 3-5, D) was downregulated as expected although in varying degrees, which indicates the persistence of some ES cells despite the differentiating environment and is to some degree expected. These results show that three days of LIF removal and then addition of RA can successfully differentiate ES cells towards different cell lineages in a reproducible way.

3.3. DNA methylation of *Oct4* during *in vitro* differentiation and its effect in the gene's expression

3.3.1. There are distinct methylation patterns in the different regulatory elements of Oct4

Methylation at the promoter of *Oct4* has been linked with its transcriptional status. Nevertheless, until very recently there was no detailed information on the establishment of the methylation pattern and how it correlates with transcription. Appreciating this gap in our knowledge, the methylation pattern of the *Oct4* 5' upstream region as it is being established during *in vitro* differentiation of E14 ES cells was analysed. Figure 3-6 shows the time course of Oct4 downregulation during the *in vitro* differentiation program used.

The results of the detailed bisulfite analysis of the entire *Oct4* upstream region as well as small part of the gene itself are shown in Figure 3-7. As Feldman *et al.* (2006) have shown for the promoter region with the lower resolution method of

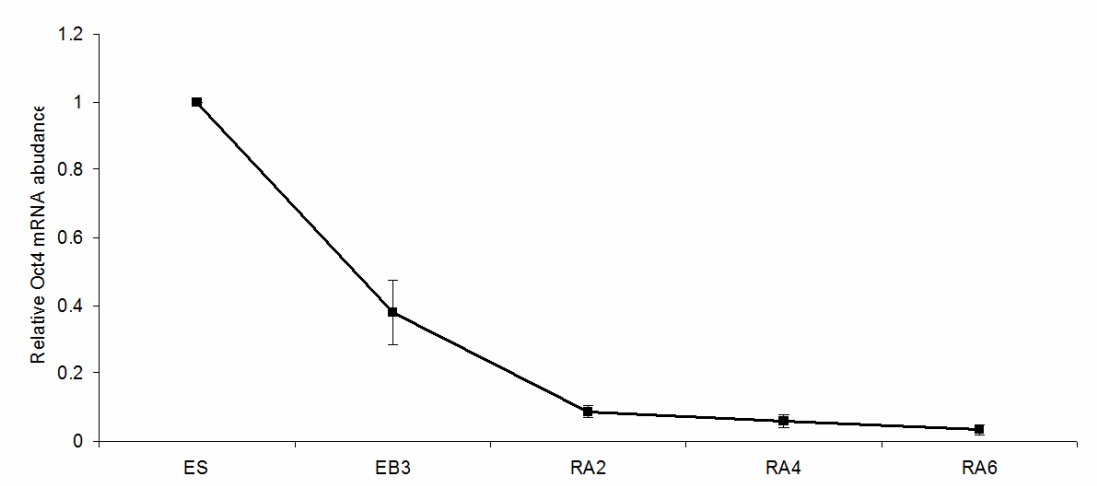

Figure 3-6. RT-qPCR expression analysis of *Oct4* in differentiating E14 ES cells. This is the average of five experiments. The error bars are the standard error of the mean for each differentiation point. The *Oct4* levels were normalised against *gapdh* and are represented here as fractions of the levels in ES cells.

methylation-sensitive PCR analysis, establishment of methylation appears to happen *after* the downregulation of transcription. In consequence, in EB3 *Oct4* levels are already decreased by half (Figure 3-6) but there is no sign of methylation in any of the examined regions (Figure 3-7). Methylation appears to start soon after RA2, when the mRNA levels of the gene have almost reached their minimum.

What has not been published previously is a thorough bisulfite analysis along the *Oct4* upstream region that includes all regulatory regions and all non-functional DNA. This information, taken together with the known distinct regulatory functions of the DE, PE and P in *Oct4* expression, could provide invaluable information on whether methylation is specific to the regions that play a role in the gene's expression in ES cells, or happens indiscriminately throughout the region.

In order to perform the analysis in each individual regulatory and non-regulatory element, the entire region was divided into seven segments as shown in Figure 3-7. Segment A contained the DE, segment C the PE, segment B was the region between the two enhancers and segment was E the region between the promoter and the PE. Finally, the promoter was subdivided into two segments; segment G contained the first few CpGs of the transcribed region and the Sp1/3-RARE site and segment F that contained the part of the promoter that has not been

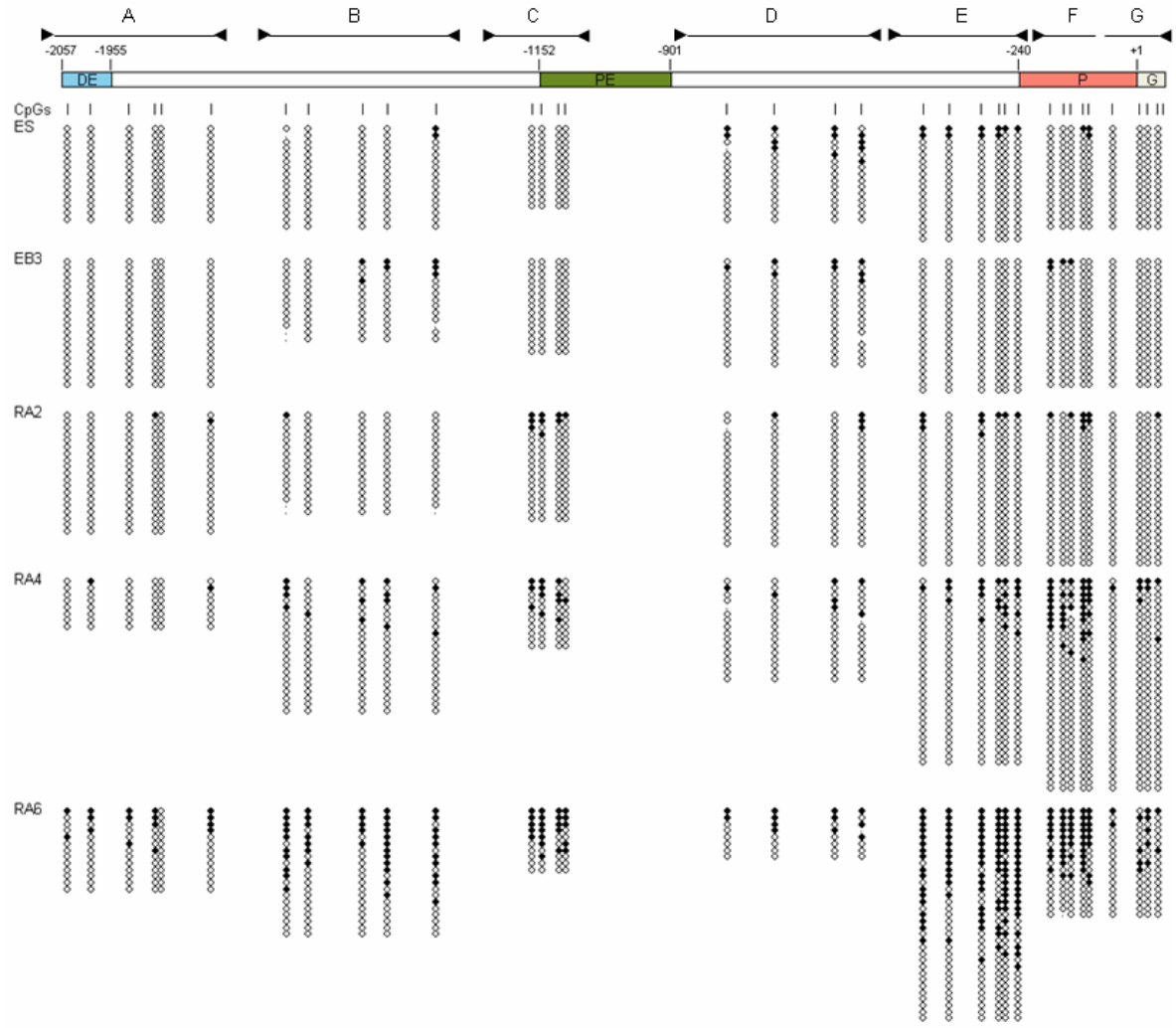

Figure 3-7. Bisulfite analysis of the entire upstream region of *Oct4* during *in vitro* differentiation of E14 ES cells. The arrowheads show the positions and direction of the primers used for walking across the region. A, B, C, D, E, F and G refer to the segments used for the statistical analysis in Figure 3-8. The positions of the regulatory elements in reference to the transcription start site are shown. DE, distal enhancer; PE, proximal enhancer; P, promoter; G, gene. Filled circles represent methylated CpGs and empty ones non-methylated CpGs.

shown to include any transcription factor binding sites. The methylation frequency in each of these segments was calculated for RA4 and RA6 (Figure 3-8). These two differentiation points were selected for the analysis to represent the initiation and advanced stage of *de novo* methylation respectively. These methylation frequencies were then subjected to the Wilcoxon two sample rank test for the null hypothesis that

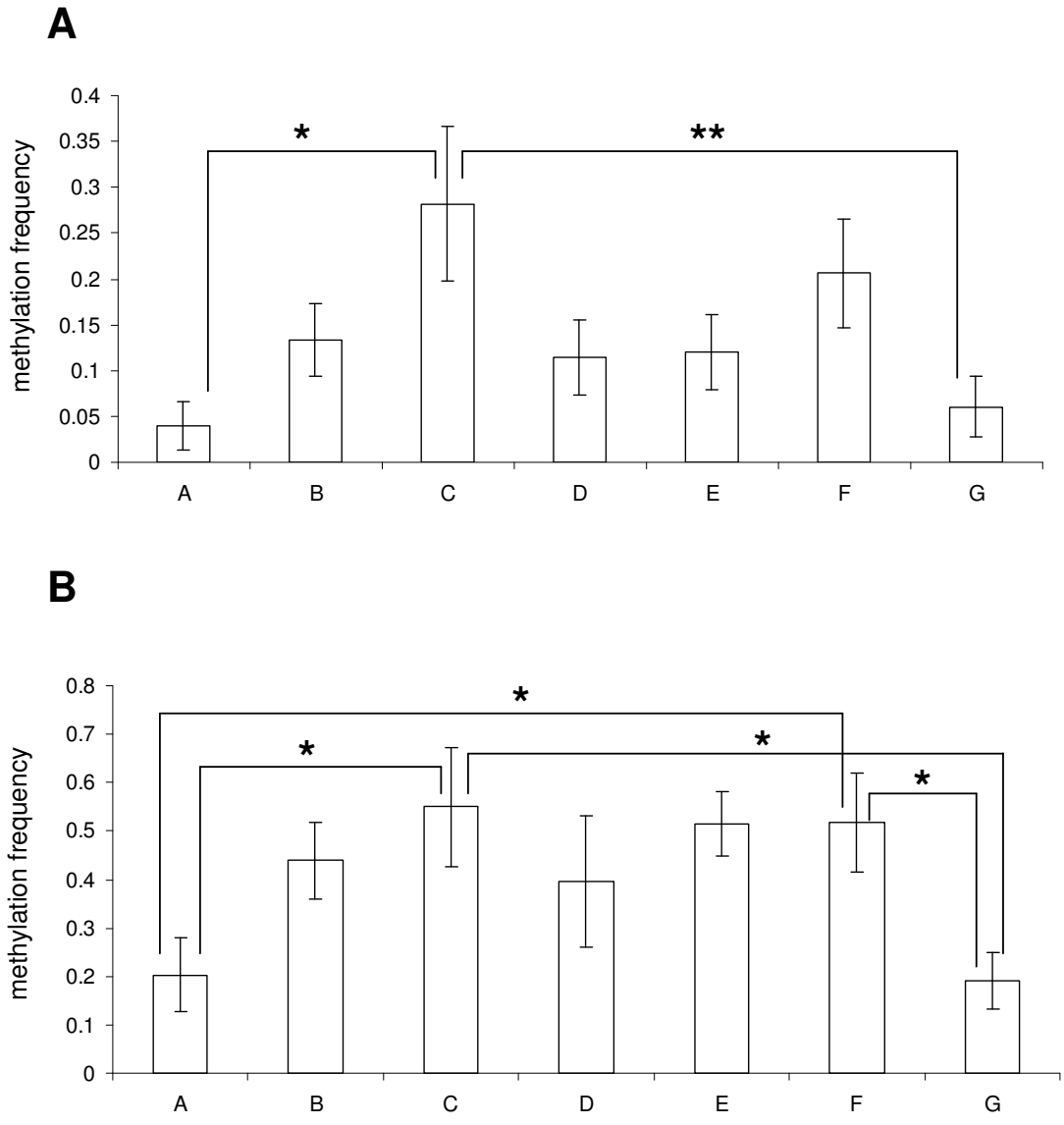

Figure 3-8. Methylation frequency for each element of the *Oct4* upstream region in the course of *in vitro* differentiation of E14 ES cells. The methylation frequency was calculated for RA4 (A) and RA6 (B) from the bisulfite data in Figure 3-7. The frequencies were calculated as number of methyl-CpGs over the total number of CpGs in each segment. The data were analysed pairwise with the Wilcoxon two sample rank test, * p≤0.05 and ** p≤0.01. The error bars show the standard error of the mean as calculated for the among clone variation per segment.

the methylation is non-uniform among the examined segments. Comparable results were obtained with the Kolmogorov-Smirnov test.

At the initiation of *de novo* methylation (Figure 3-8, A), the PE, which has a role in *Oct4* expression in the epiblast and has been shown to lose LRH-1 occupancy when repressed, appeared to be preferentially methylated in comparison to the DE and the Sp1/3-RARE containing segments. The first CpG upstream of the transcription start position (Figure 3-7) was virtually devoid of methylation, something that has also been observed by Gu *et al.* (2006). Methylation between consecutive elements did not show a significant variation, indicating that methylation might be established in specific regions and then "leak" sideways. As the establishment of the methylation pattern progressed (Figure 3-8, B) methylation increased throughout the upstream region of *Oct4*. However, the DE and the Sp1/3-RARE containing segments (A and G) continued to have very low methylation levels. At this stage, the F fragment of the promoter region also had significantly higher methylation levels than the more upstream G fragment and the DE. The fragments that do not contain any regulatory regions appeared to be following the methylation pattern of their surroundings.

3.3.2. DNMT3a is the main de novo DNA metyltransferase present at the time Oct4 methylation is being established

Although it is now generally accepted that DNA methylation at the promoter of *Oct4* appears after downregulation of the gene, there has been some controversy regarding the role of DNMT3a and b. Experiments with DNMT3a/b -/- cells have been contradictory (Jackson *et al.* 2004; Feldman *et al.* 2006; Watanabe *et al.* 2002). In order to explore the importance of the DNA methylatransferases for the downregulation of *Oct4* in the present system, the efficiency of *Oct4* downregulation in differentiating DNMT3a/b -/- and DNMT1 -/- ES cells (passage number 20 at the beginning of *in* vitro differentiation) was investigated (Figure 3-9). The downregulation of *Oct4* in these cells was indistinguishable from the wild-type (Figure 3-6), indicating that DNMTs –like DNA methylation– are dispensable for the gene's silencing.

Despite the fact that DNMTs seem to be dispensable for *Oct4* downregulation, *de novo* methylation does take place in wild-type cells and there is

substantial evidence that GCNF and G9a recruit *de novo* DNA methyltransferases albeit indirectly. It is not clear though whether the recruitment involves DNMT3a, DNMT3b or both. Watanabe *et al.* (2002) have shown that the levels of DNMT3a and b change during embryonic development. In order to investigate whether the DNMT3a and b levels remain constant during the *in vitro* differentiation program, the mRNA level in each stage were measured by RT-qPCR (Figure 3-10). Both the DNMT3 enzymes start with similar mRNA levels relative to *gapdh* in ES cells. Nevertheless the level of DNMT3b falls by about 90% and by stage RA4 it is significantly lower than DNMT3a. This means that at the time when methylation of the *Oct4* upstream region takes place, the main DNA methyltranferase present to carry out the *de novo* methylation is DNMT3a.

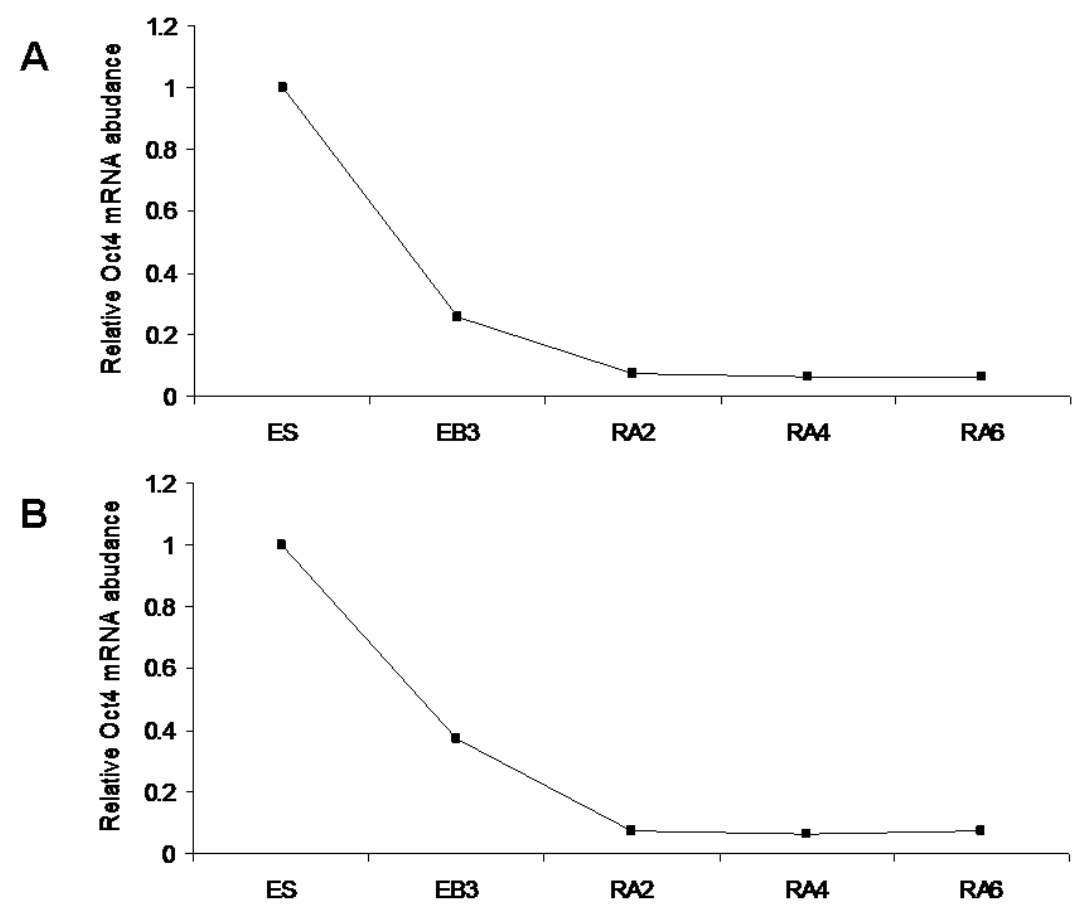

Figure 3-9. RT-qPCR expression analysis of *Oct4* in differentiating DNMT3a/b -/- (A) and DNMT1 -/- (B) ES cells. The mRNA levels were normalised against *gapdh* and are shown here relative to the levels in ES cells.

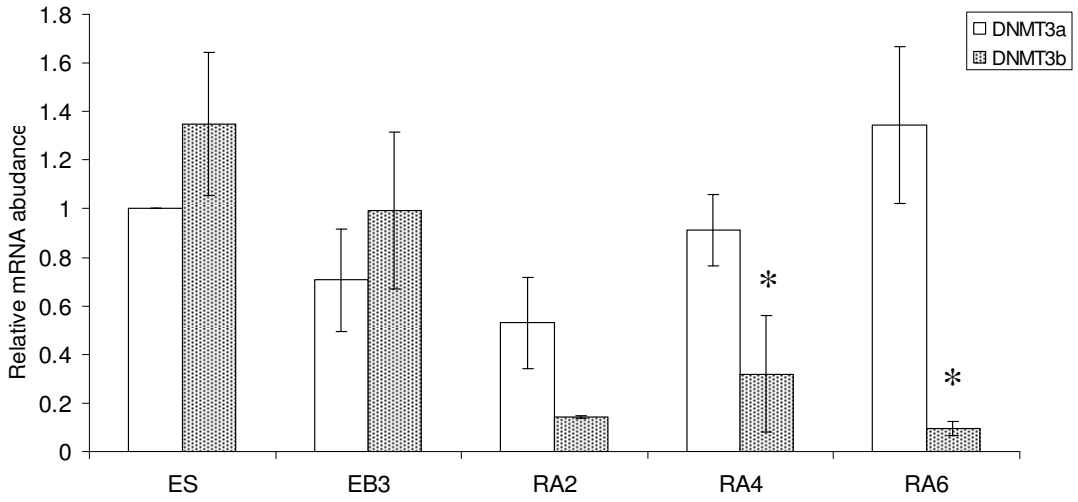

Figure 3-10. RT-qPCR analysis of DNMT3a and DNMT3b in differentiating E14 ES cells. This is the average of three experiments.The mRNA levels were normalised against *gapdh* and are shown here relative to the DNMT3a level in ES cells. The bars are the standard error of the mean for each differentiation point. Stars indicate significantly ($p \leq 0.05$) lower levels of DNMT3b in comparison to DNMT3a (paired Student's t-test).

3.4. Histone modification changes of the Oct4 distal enhancer during differentiation

Since methylation levels of the distal enhancer were comparable to those of the promoter (Figure 3-7), it was further investigated if this similarity extends to the histone modification changes. The promoter has been shown to lose H3K4 methylation and H3 acetylation (Feldman *et al.* 2006), while H4 acetylation is reduced throughout the upstream region during differentiation (McCool *et al*. 2007). Analysis of the DE (Figure 3-11) showed that H3 acetylation was highly enriched after differentiation and H4 acetylation appeared depleted. The level of trimethylH3K4, which is typical of active chromatin regions –like H3 and H4 acetylation– did not show any change between the pluripotent and differentiated cells although it has been shown to decrease at the promoter. The same result has been observed by Aoto *et al.* (2006).

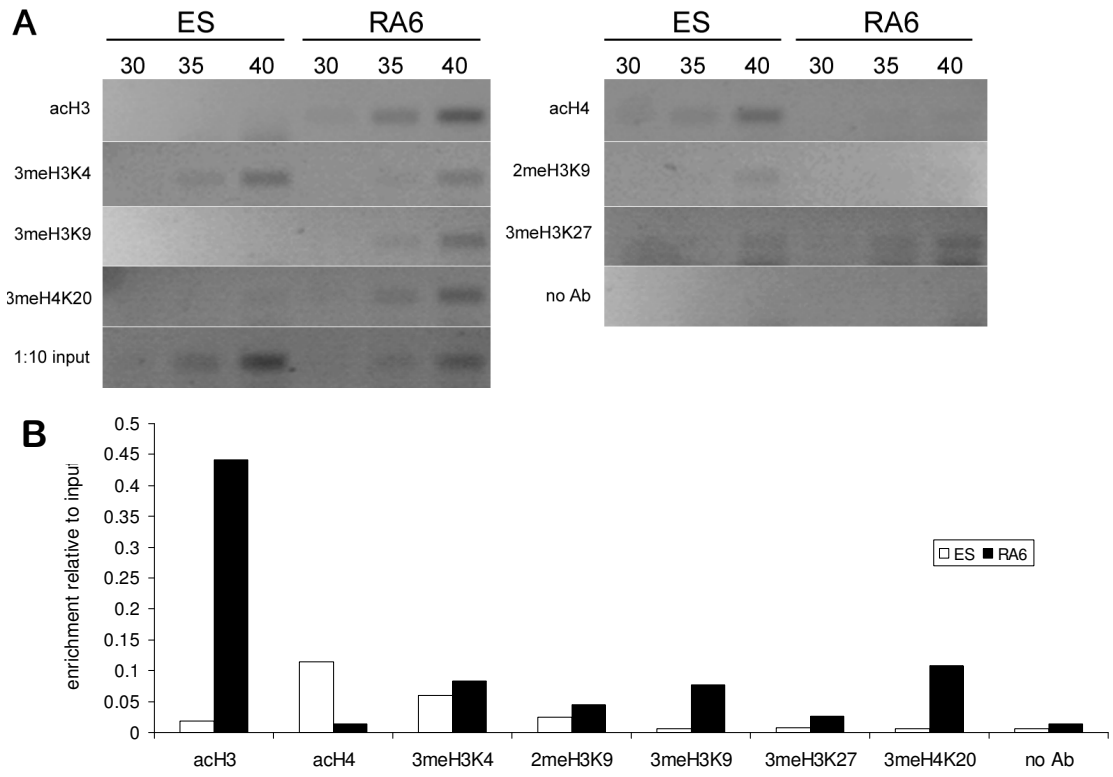

Figure 3-11. Chromatin immunoprecipitation of histone modifications in the DE element of Oct4 before (ES) and after (RA6) differentiation of E14 cells. The immunoprecipitated DNA was PCR amplified for 30, 35 and 40 cycles (A). (B) Quantification of the PCRs shown in (A) relative to input (*i.e.* the signal from the ChIPed samples divided by the signal of input times ten).

As for the "repressive" chromatin modifications, trimethylH4K20 was observed to increase at the distal enhancer in the silenced gene. This is probably related to the decrease in H4 acetylation as the two modifications have been shown to be mutually excusive (Sarg *et al.* 2004) . On the other hand, the polycomb-related trimethylH3K27 did not show any changes. The latter does not come as a surprise since Aoto *et al.* (2006) have shown that this modification is enriched only at terminally differentiated, not proliferating, *in vitro* generated neurons and not their precursors. Dimethyl H3K9 levels do not seem to change but there is a large enrichment in trimethyl H3K9 as observed by Feldman *et al.* (2006) for the promoter. All in all, in comparison to the histone modification changes studied at the promoter, trimethylH3K9 enrichment and acetylH4 depletion seem to be shared with the distal enhancer.

3.5. The effect of G9a on *Oct4* methylation.

Feldman *et al.* (2006) have shown that the histone methyltransferase, G9a is recruited at the promoter of *Oct4* and is responsible for the enrichment of trimethylH3K9 after differentiation. TrimethylH3K9 subsequently attracts the *de novo* DNA methyltransferases to the region. The observation that trimethylH3K9 levels also increase in the DE raises the possibility that G9a is also recruited to the DE. The effect that G9a depletion has directly on the methylation patterns in the silenced promoter, or any other element in the *Oct4* upstream regulatory region, was not investigated. For this reason, detailed bisulfite analysis of the *Oct4* upstream region was also conducted for the G9a -/- 2-3 ES cells. The goal of this analysis was to investigate the direct effect of G9a on the establishment of DNA methylation throughout the upstream region of *Oct4*.

These cell lines are from a different genetic background (C57BL/6) than the E14 ES cells already analysed (129/Ola) and for this reason it was first examined whether the genetic background affects the ES cell response to *in vitro* differentiation conditions.

3.5.1. ES cells of different genetic background follow a different differentiation program after induction with RA

Wild-type COL4 ES cells (C57BL/6) were induced to differentiate by removal of LIF for three days and then addition of RA as described (section 3.2.1). The analysis of the differentiation process was performed with various differentiation markers as before (section 3.2.2) and is shown in Figure 3-12. Although there are many similarities in the expression of the various developmental markers with the E14 ES cells (Figure 3-5) there are also some striking differences. The most prominent difference is the expression of the mesodermal markers in this cell line (Figure 3-12, B), something that was not observed on E14 cells. Although the endodermal and ectodermal markers *TTR* and *Msx3* respectively are not expressed in this cell line, the other markers of these lineages however show very similar

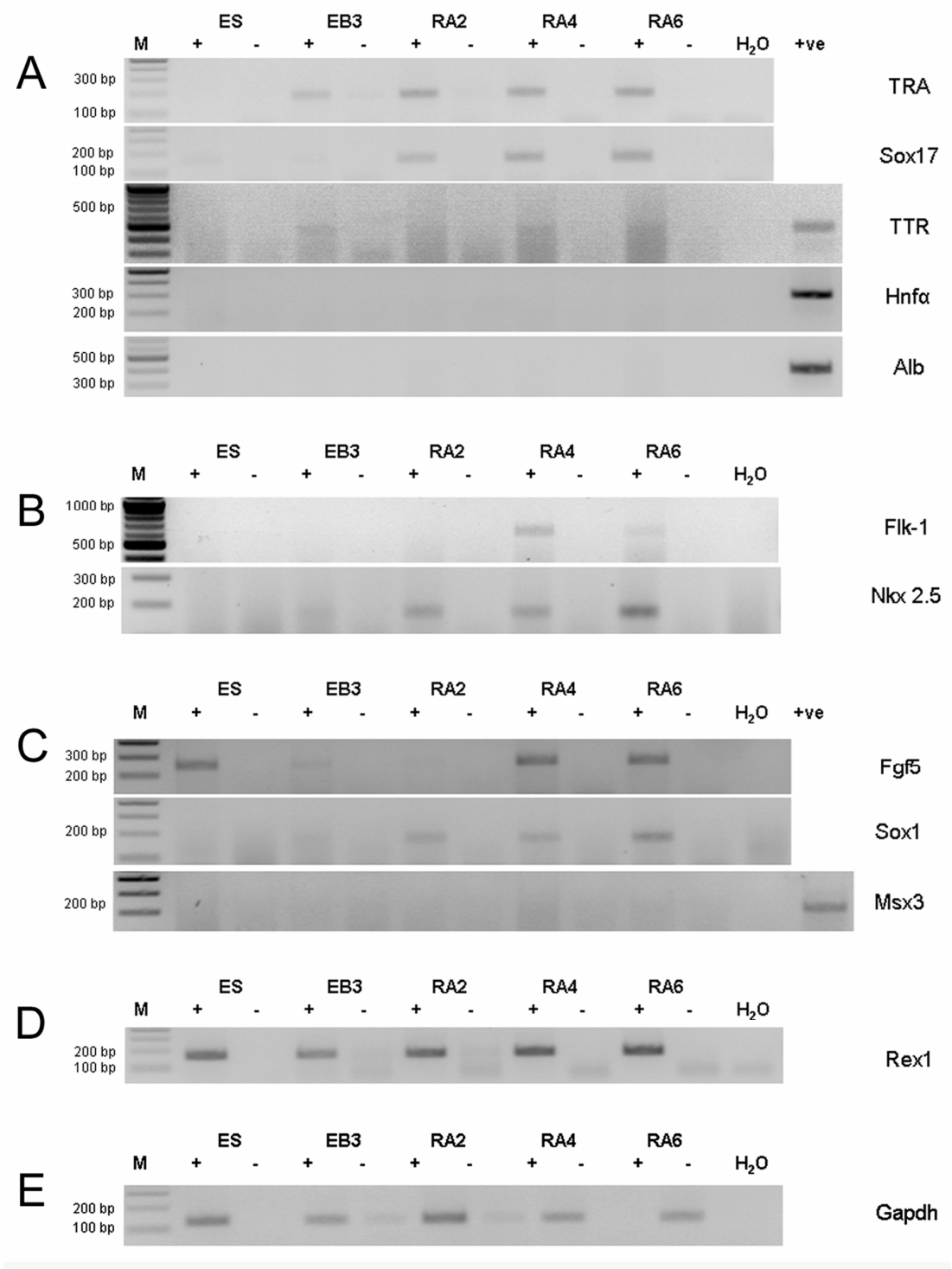

Figure 3-12. Expression of developmental markers during the course of in vitro differentiation of COL4 ES cells. + : reverse transcribed RNA, – : mock-reverse transcribed RNA, +ve: positive control. M is the molecular weight marker. (A) Endodermal markers. (B) Mesodermal markers. (C) Ectodermal markers. (D) Stem cell marker. (E) Loading control. Refer to Figure 3-5 for details.

expression patterns to the E14 ES cells. Surprisingly, despite the successful upregulation of various differentiation markers, the ES cells marker *Rex1* (Figure 3-12, D) is not downregulated.

In summary, the ES cell lines of the two genetic backgrounds show differences in the direction they differentiate to in the presence of RA. The 129/Ola cell line shows inability to differentiate towards mesodermal lineages, whereas the C57BL/6 cells can. Moreover, although both cell lines expressed ectodermal markers, the precise markers detected varied between cell lines indicating that different ectodermal directions were followed. Finally, the C57BL/6 cells showed complete inability to downregulate the ES cell marker *Rex 1* in the given differentiation time.

3.5.2. ES cells of different genetic backgrounds show differences in the establishment of methylation

Having shown that the wild-type ES cells of different genetic background do not behave identically under differentiation conditions, the wild-type methylation patterns were re-analysed in differentiating COL4 ES cells. The results of these analyses are shown in Figure 3-13 (B) and Figure 3-14. There is variation in the establishment of methylation in the two different genetic backgrounds; E14 ES cells (Figure 3-7) accumulate some methylation in the DE at the end of the differentiation program, while COL4 fail to gain any methylation at this region ($p=0.05$). Methylation levels of the two wild-type cell lines at the other regions examined show no difference.

3.5.3. G9a -/- ES cells fail to establish distinct methylation patterns in the regulatory elements of Oct4

Examination of the effect that the absence of G9a has on the establishment of methylation at the *Oct4* upstream regulatory region revealed that the mechanism is indeed impaired throughout the upstream region and not only at the promoter (Figure 3-13C and Figure 3-14). Initially, in RA4, methylation in G9a -/- ES cells increased, and the methylation pattern was indistinguishable to the wild-type cells. The establishment of DNA methylation in the *Oct4* upstream region appeared to start from the PE (segment C), like in the E14 cells (Figure 3-7). This methylation

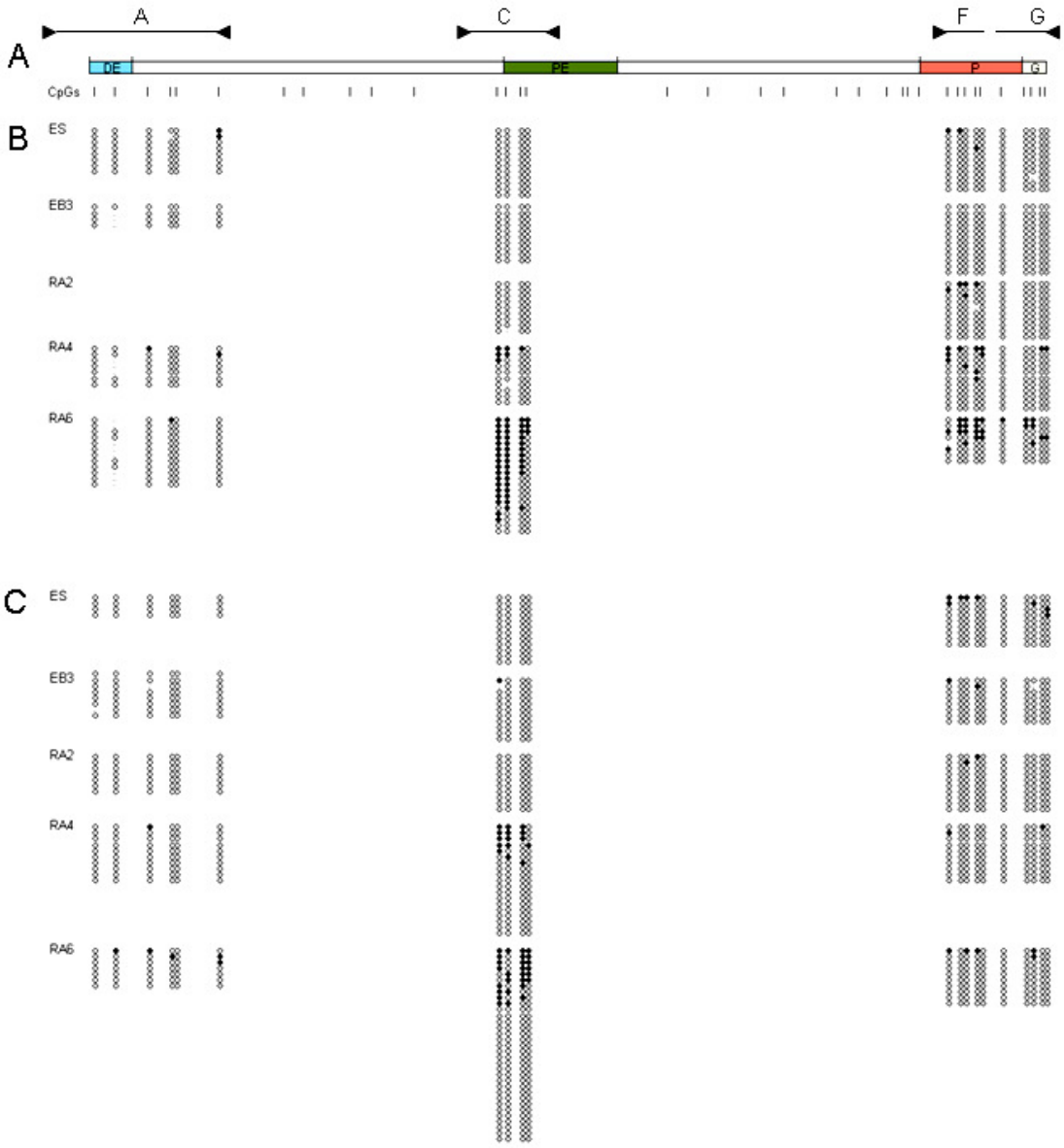

Figure 3-13. Bisulfite analysis of the DE, PE and P of Oct4 during in vitro differentiation of COL4 and 2-3 ES cells. (A) Map of the upstream Oct4 region. The arrowheads show the positions and direction of the primers used. A, C, F and G refer to the segments used for the statistical analysis in Figure 3-14. DE, distal enhancer; PE, proximal enhancer; P, promoter; G, gene. (B) Wild-type COL4 ES cells. (C) G9a -/- 2-3 ES cells. Filled circles represent methylated CpGs and empty ones non-methylated CpGs.

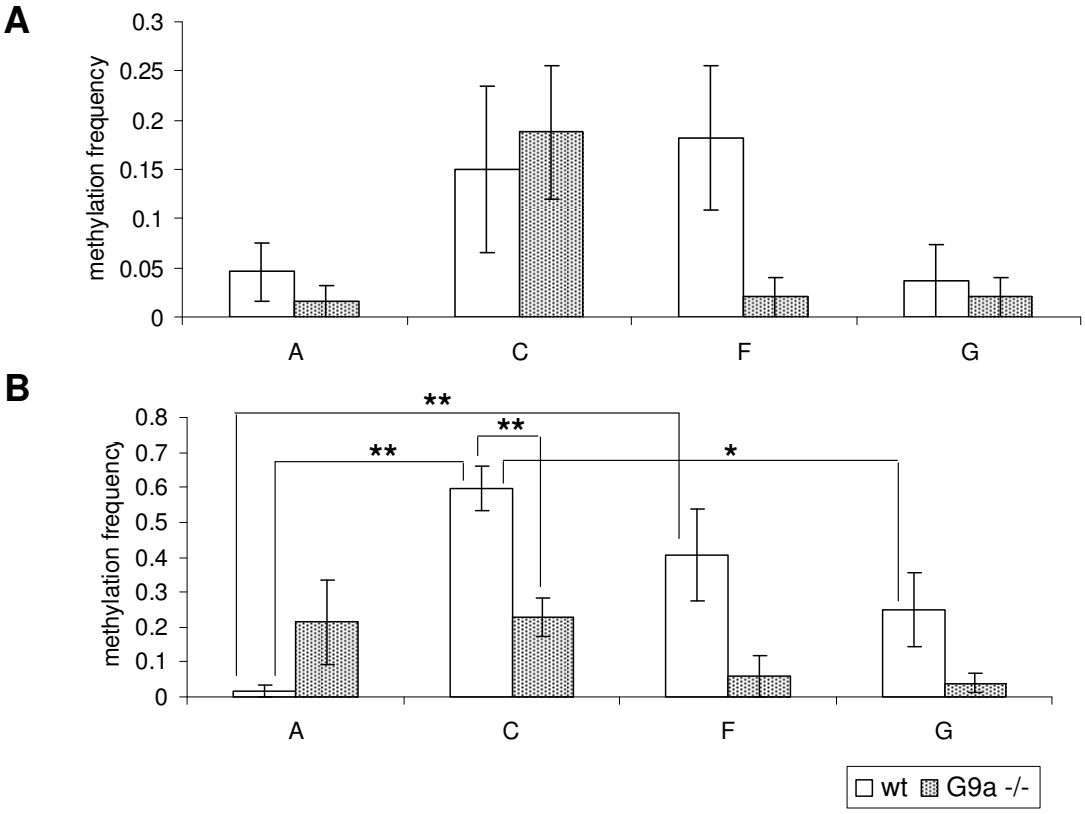

Figure 3-14. Methylation frequency for the DE, PE and P elements of the *Oct4* upstream region in the course of *in vitro* differentiation of G9a -/- 2-3 ES cells and their wild-type controls (COL4). The methylation frequency was calculated for RA4 (A) and RA6 (B) from the bisulfite data in Figure 3-13. The frequencies were calculated as number of methyl-CpGs over the total number of CpGs in each segment. The data were analysed pairwise with the Wilcoxon two sample rank test, * p≤0.5 and ** p≤0.1. The error bars show the standard error of the mean as calculated for the among clone variation per segment.

however failed to continue as it did in the wild-type COL4 cells. By the end of the differentiation protocol, there was a statistically significant difference in the methylation level of the PE (segment C) between the G9a -/- and wild-type ES cells. Examination of the promoter region (segments F and G), revealed that methylation there was also very low. It would appear from this study that initiation of methylation is unaffected in the G9a -/- ES cells and the methylation levels at the PE have correctly risen by RA4. However, establishment of high methylation levels at the PE –and to a less extent to the DE and P– at a later stage is much impaired. This difference has been missed in the previous studies.

3.5.4. *Oct4* mRNA downregulation is not impeded by the absence of G9a

It has been reported that *Oct4* downregulation is not affected in the short term by the absence of G9a (Feldman *et al.* 2006). Examination of whether this is the case in this differentiation system confirmed that G9a -/- cells downregulate Oct4 expression albeit more gradually than the wild-type cells (Figure 3-15, A). It is

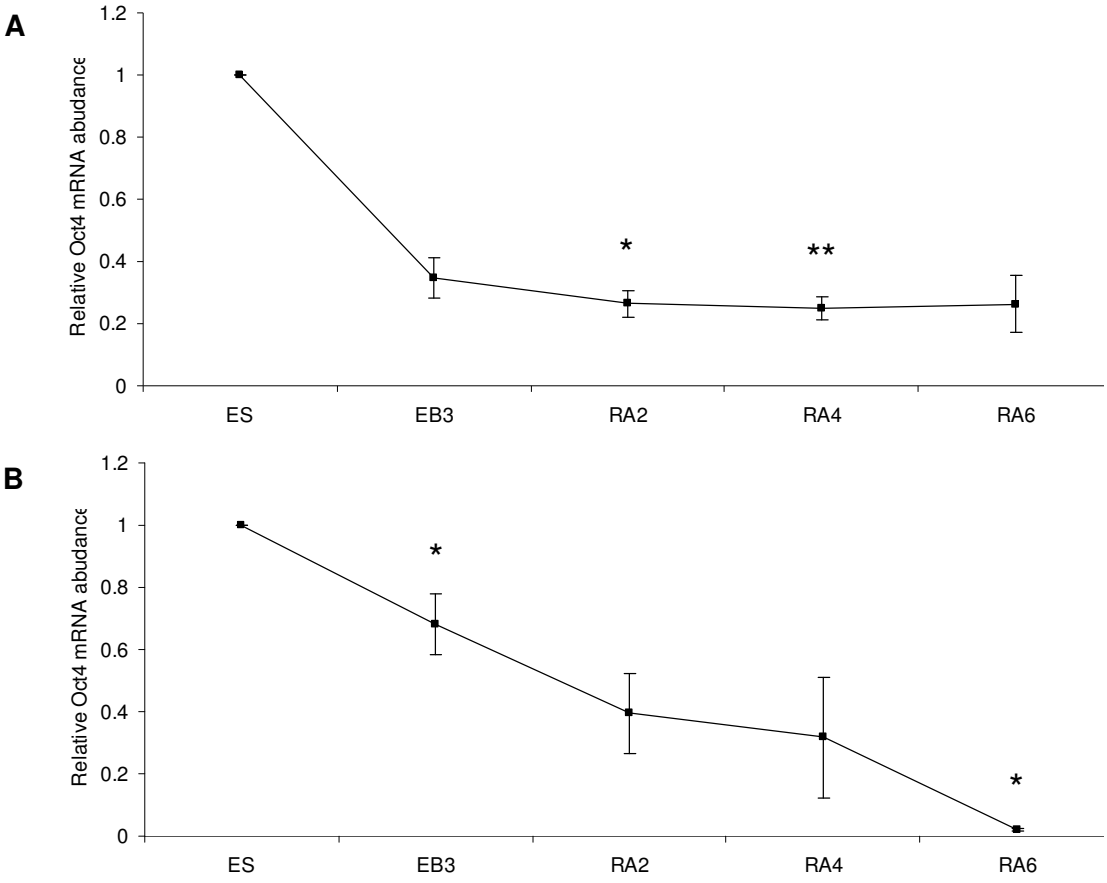

Figure 3-15. Expression analysis of *Oct4* in COL4 (A) and 2-3 (B) ES cells using RT-qPCR. Each plot shows the average of three experiments. The error bars are the standard error of the mean for each differentiation point. The *Oct4* levels were normalised against *gapdh* and are represented here as fractions of the levels in ES cells. * p≤0.05 and ** p≤0.01 (Student's t-test). (A) Wild-type COL4 ES cells. The asterisks indicate significant difference from the equivalent differentiation stage of E14 ES cells (Figure 3-16). (B) G9a -/- 2-3 ES cells. The asterisks indicate significant difference from the equivalent differentiation stage of COL4 ES cells.

interesting that the wild-type COL4 ES cells show inability to completely shut down the gene, at least to the levels of the E14 cell line (Figure 3-15, B). This observation could explain the previously observed failure to downregulate Rex1 in this cell line (Figure 3-12), as Oct4 is this gene's direct activator (Ben-Shushan *et al.* 1998)

3.6. *Nanog* levels mirror *Oct4* transcription fluctuations in different cell lines

The observation that *Oct4* downregulation follows different patterns in the different cell lines examined (E14, COL4 and 2-3) raises questions about the reason for such a variation. The methylation analyses have shown that there are different methylation patterns at the three regulatory elements of *Oct4* in these cell lines. This however, can not explain this difference since in all cases methylation appears second. An explanation could come from the effect of some transcription factor. As a possible candidate, the expression pattern of *Nanog* in these cell lines is examined. *Nanog* is successfully downregulated, as expected, in the differentiating E14 and 2-3 ES cells that can also downregulate *Oct4* expression to basal levels (Figure 3-16, A). In contrast, COL4 cells that were unable to completely shut down *Oct4,* fail to downregulate *Nanog* (Figure 3-16, B). This poses the egg and hen paradox: is failure to repress *Oct4* causing *Nanog* levels to rise or *Nanog* is not repressed because Oct4 is still present? The answer is probably both, and this is a forceful demonstration of the power that feedback mechanisms have in biological systems.

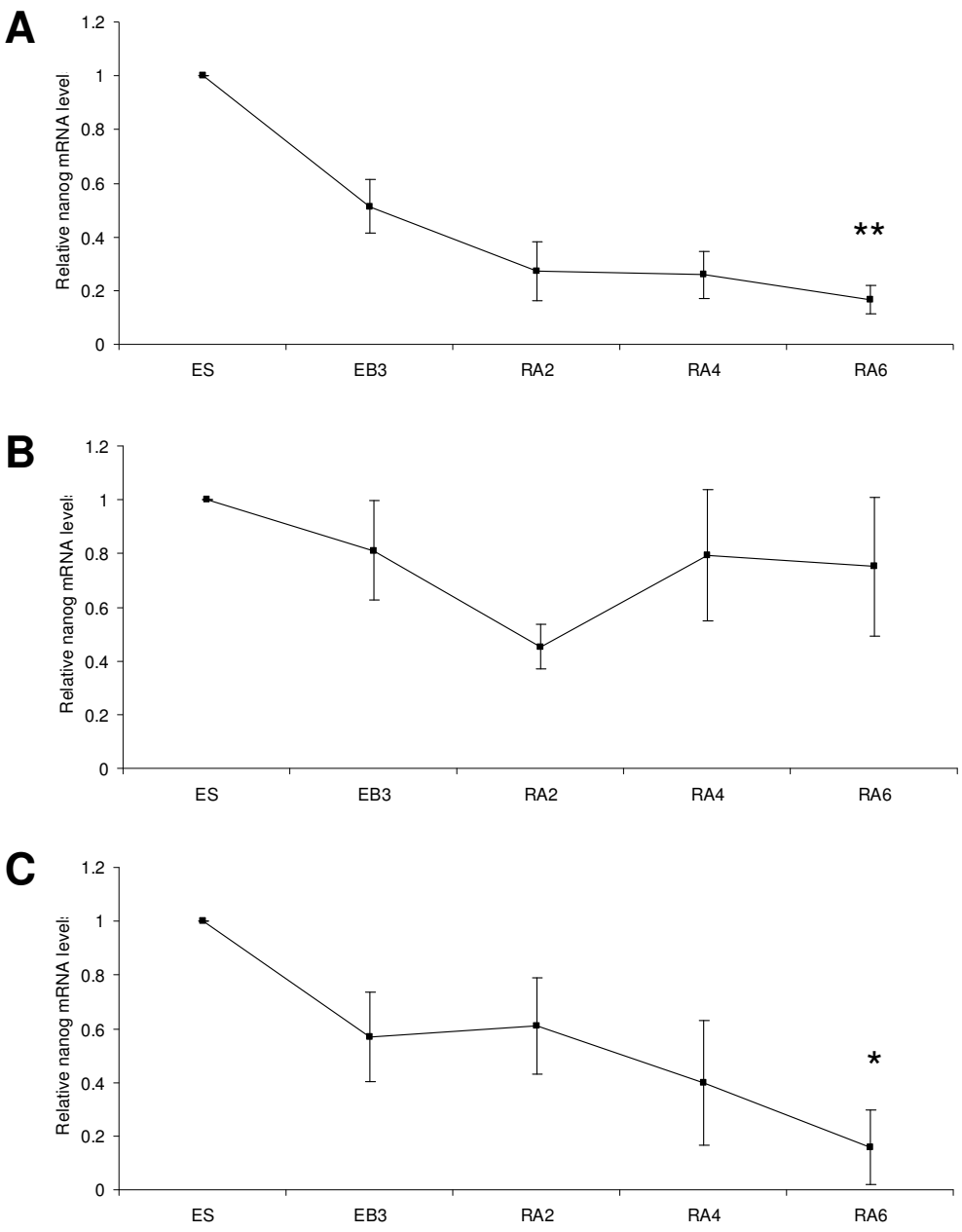

Figure 3-16. Expression analysis of *nanog* in COL4 and 2-3 ES cells using RT-qPCR. The error bars are the standard error of the mean for each differentiation point. The *Oct4* levels were normalised against *gapdh* and are represented here as fractions of the levels in ES cells. The asterisks indicate significant difference from the mRNA levels in the ES cells of the same cell line.* $p \leq 0.05$ and ** $p \leq 0.01$ (Student's t-test). (A) Wild-type E14 ES cells. The average of five experiments is shown. (B) Wild-type COL4 ES cells. The average of three experiments is shown and the asterisks indicate significant difference from the E14 ES cells. (C) G9a -/- 2-3 ES cells. The average of three experiments is shown and the asterisks indicate significant difference from the COL4 ES cells.

3.7. Discussion

Oct4 is the best studied pluripotency transcription factor with respect to its regulation. Intense research in the last years provides evidence for a regulation that involves a complex network of activators, suppressors, feedback loops and epigenetic changes that act on the promoter and the two enhancers of the gene. Although many of these players are known, many mediators of interactions still remain to be discovered. More importantly, the orchestration of all the factors at the gene's three regulatory elements in order to stop transcription as a response to the differentiation stimulus is still unknown. Finally, most emphasis until now has been given to the promoter region and our knowledge about the events that take place on the two enhancers is very limited. This study investigated the link between transcriptional repression and methylation of the gene's entire regulatory region.

Assesment of the *in vitro* differentiation protocol as a means to downregulate *Oct4*

In vivo, *Oct4* is silenced in the trophoblast and as the epiblast differentiates into the three main cell lineages. Assuming that there is only one mechanism for the gene's downregulation, *in vitro* differentiation of ES cells into embryoid bodies has been used for its study (Sato *et al.* 2006; Gu *et al.* 2006; Feldman *et al.* 2006 and this study). In all the previous studies, ES cell differentiation was induced by simultaneous removal of LIF and addition of RA. In the present study LIF was first removed for three days and then RA was added to the medium for another six days. Comparison of the timing of the gene's expression and methylation between the present and previous studies is difficult as there appears to be variation between different laboratories; expression of *Oct4* is reported to reach its minimum after between two and six days of differentiation, while DNA methylation seems to first appear after between one and six days. Despite the variability, the results of the present study fall within the time ranges described in the previous reports; expression of *Oct4* reaches its minimum two days after addition of RA, while DNA methylation first appears four days after addition of RA. What is more important, in all the studies, DNA methylation appears after the gene is downregulated.

It should be noted that in the present study (Figure 3-6) as well as in other reports (Sato *et al.* 2006; Gu *et al.* 2006), *Oct4* expression in the EBs, as assessed by RT-PCR, never reaches zero but there is always some residual, very low expression of the gene. This could be attributed to the presence of a small, resistant population of ES cells that are not differentiated. Such an assumption is supported by the residual expression of *Rex1* (Figure 3-5, D) and *nanog* (Figure 3-16, A) too. Alternatively to this hypothesis, *Oct4* expression in all cells could approach zero asymptotically. If this is the case, then Figure 3-6 shows the kinetics of *Oct4* downregulation. The simplest way to test experimentally which of the two hypotheses is correct, whould be to stain the EBs of various differentiation stages with antibodies against *Oct4* and/or some other ES cell marker such as SSEA-1. If the first hypothesis of the presence of a residual population of ES cells is correct, then an increasing population of cells should be negative for the staining at each differentiation stage but Oct4-positive cells should persist. If on the other hand, *Oct4* downregulation happens gradually in all the cells, the intensity of the staining in each cell should decrease in each differentiation stage. Similarily, EGFP or luciferase expression, driven by the *Oct4* regulatory elements in differentiating ES cells could be used to monitor the kinetics by which the gene is downregulated.

Patterns of methylation establishment in *Oct4*

It has recently been reported that the downregulation of *Oct4* does not depend on DNA methylation although methylation does invariably appear at a later stage (Feldman *et al.* 2006; Gu *et al.* 2006; Sato *et al.* 2006). The present results support this observation and further show that there seems to be a "DNA methylation wave" that does not extend uniformly along the gene's upstream region, but it rather happens preferentially at some elements and not others (Figure 3-17). In more detail, methylation is first detected at the proximal enhancer. As differentiation proceeds, DNA methylation seems to increase and also spread towards the promoter and the distal enhancer.

There are two possible explanations for the observed pattern of methylation establishment at the *Oct4* upstream region; methylation is targeted first to the proximal enhancer through specific recruitment of DNMT3s, or the proximal enhancer becomes methylated first because it is not protected by transcription

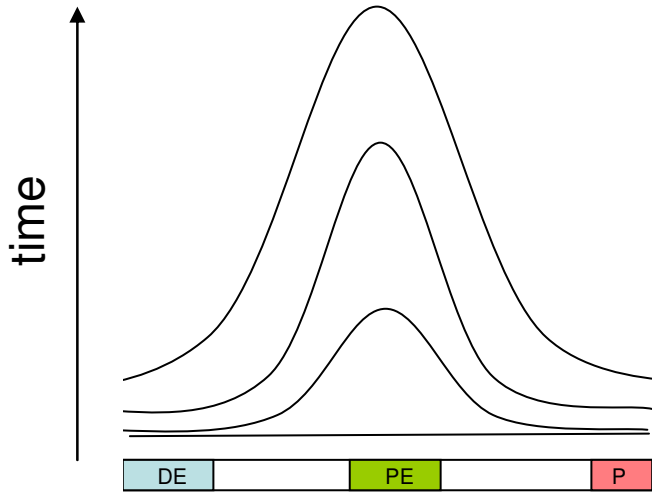

Figure 3-17. Schematic diagram of the methylation establishment pattern at the *Oct4* upstream region. The diagram models the observed methylation levels at each element of the *Oct4* upstream region with time, after induction of differentiation. DE, distal enhancer; PE, proximal enhancer; P, promoter (segment G, Figure 3-7 and Figure 3-13)

factors. A definitive answer to the question of which of the two mechanisms is responsible for the observed pattern of methylation establishment would come from ChIP mapping of DNMT3s along the *Oct4* upstream region. In the absence of such data, the current study together with previous footprinting analyses of the regulatory elements, favour the passive mechanism of methylation establishment at the proximal enhancer. In more detail, no suppressors are known to bind to the proximal enhancer and *in vivo* footprinting has shown that this region is protected in undifferentiated EC and ES cells but the occupancy is lost upon treatment with RA (Minucci *et al.* 1996; Okazawa *et al.* 1991). The promoter on the other hand, is resistant to methylation only in part; the portion of the promoter that resists methylation is the same that contains recognition elements for the repressor GCNF and the activator SF-1 and has been shown to be protected in both undifferentiated and differentiated EC cells (Pikarsky *et al.* 1994). The central CpG of the potential Sp1 binding site, in particular, remains virtually free of methylation in this (Figure 3-7and Figure 3-13) and other studies (Gu *et al.* 2006). The rest of the promoter, where no transcription factors are known to bind, is not protected in footprinting assays and it gains significant levels of methylation, comparable to those of the

proximal enhancer. The footprinting data, together with the sharp, significant drop of DNA methylation levels between these two adjacent regions of the promoter (Figure 3-8) testify for a mechanism in which protection by transcription factor binding impedes DNA methylation establishment. Such a mechanism has been reported before for the mouse *Aprt* gene in which Sp1 binding is protecting the gene's CpG island from methylation (Brandeis *et al.* 1994; Macleod *et al.* 1994; Mummaneni *et al.* 1998). In an analogous manner, lack of protection by transcription factor binding at the proximal enhancer, would make this element vulnerable to the action of DNA methyltransferases that have been recruited at the promoter. These data support the hypothesis that methylation at the proximal enhancer and the non-regulatory sequences that surround it, is not targeted but rather happens where the opportunity arises, as transcription factors are removed and a region becomes exposed. As it will be discussed extensively later, DNMT3 recruitment to the promoter could have a function that is not mediated by *de novo* DNA methylation.

If this is indeed the case, then the question arises: why is methylation first established at the proximal enhancer and not throughout the central part of the *Oct4* upstream region (segments B through E, Figure 3-7)? The distal enhancer, but not the proximal enhancer, is involved in *Oct4* expression in ES cells (Yeom *et al.* 1996). Like at the promoter region, a significant delay in methylation was also observed in the distanl enhancer (Figure 3-7). Moreover, after induction of repression, similar changes in the levels of H3K9 methylation and H4 acetylation were observed for the distal enhancer (Figure 3-11) as for the promoter (Feldman *et al.* 2006). Finally, low resolution analysis of the histone modifications in ES cells in the 40 Kb region surrounding *Oct4*, gives evidence that H3/4 acetylation, H3K4 methylation and H3S10 phosphorylation marks are shared between the promoter and distal enhancer but not the promoter and the proximal enhancer (Aoto *et al.* 2006). These data suggest that the same factors are acting on the promoter and distant enhancer –but not on the proximal enhancer– despite the fact that the two elements are separated by more than 2Kb. A simple model of enhancer action from a distance is DNA bending and formation of a loop that brings the enhancer closer to the gene (Li *et al.* 2006b). In the *Oct4* upstream region, this would mean that the distal enhancer and the promoter are brought in close proximity, excluding the proximal enhancer from the

regulation of the gene's transcription in ES cells. This in turn would explain why the proximal enhancer acquires methylation first, as this element is located midway between the promoter and the distal enhancer (approximately 1Kb from each), thus has the biggest distance from the regions that are occupied by transcription factors and potentially protected from DNA methylation.

Evidence for DNMT3a specificity on *Oct4*

According to the expression profile of DNMT3a and b during the course of differentiation (Figure 3-10), the main *de novo* DNA methyltransferase present at the time of methylation establishment in the upstream region of *Oct4* is DNMT3a. This suggests that DNMT3a alone could be responsible for the *de novo* methylation of *Oct4*. More support for this comes from experiments in which HP1β, a known partner of DNMT3a, and DNMT3a itself were immunoprecipitated on the *Oct4* promoter (Feldman *et al.* 2006). Furthermore, co-transfection of GCNF and DNMT3a caused methylation of the *Oct4* promoter in ES cells (Sato *et al.* 2006). Nevertheless, GCNF co-immunoprecipitates with both DNMT3a and b *in vitro* (Sato *et al.* 2006) and there are no published negative immunoprecipitation results that show clearly that DNMT3b can not be found at *Oct4* .The absence of such information may be due to the low levels of DNMT3b. The observation that DNMT3a operates in a distributive fashion while DNMT3b is processive (Gowher and Jeltsch 2002; 2001) agrees with DNMT3a being the main DNA methyltransferase acting on *Oct4*; the processive DNMT3b does not allow intermediate, partially-methylated products to populate, while the distributive DNMT3a does. It was the latter situation that was observed in *Oct4* (Figure 3-7 and Figure 3-13). It has also been shown that DNMT3a is very inefficient in methylating DNA assembled in nucleosomes and it has much greater activity on naked DNA (Takeshima *et al.* 2006). This, in turn, re-enforces the idea that exposed DNA becomes methylated during silencing of *Oct4*. The specificity of DNMT3a on *Oct4* could be tested more directly by examining this region's methylation levels in ES cells that were differentiated in the absence of either DNMT3a or DNMT3b. If DNMT3a is indeed specific for *Oct4*, then methylation should not be established in its absence and not affected by the absence of DNMT3b.

Significance of DNA methylation establishment at *Oct4*

It is hard to imagine that, if *de novo* DNA methyltransferases are recruited to the promoter in the course of *Oct4* silencing as the experimental evidence suggest, the cell would invest energy in such interactions if there were no need for them at the locus. Nevertheless, expression analysis of *Oct4* in differentiating DNMT3a/b -/- ES cells here (Figure 3-9) and elsewhere (Gu *et al.* 2006) has shown that the gene can be successfully silenced even in the absence of DNMT3s. On the other hand Jackson *et al.* (2004) have observed that *Oct4* silencing is impeded in these cells. In the latter study, the null cells used were of advanced passage number and were severely hypomethylated, while in this study, the cells were of relatively low passage number and were expected to maintain some level of DNA methylation. It cannot be excluded that the difference between the two studies is because of secondary effects of the global levels of DNA methylation.Alternatively, the decisive difference between these two opposite observations could be the *in vitro* differentiation protocol used. In this study, as well as in the study by Gu *et al.* (2006), differentiation was induced with RA, while Jackson *et al.* (2004) differentiated with simple LIF removal. As seen in Figure 3-3 and Figure 3-4 there is a significant delay on *Oct4* downregulation and methylation when RA is not included in the differentiation protocol. This delay is also apparent in the necessary for downregulation timescale in the wild-type cells in the work of Jackson *et al.* (2004) that lasted several weeks and that of Gu *et al.* (2006) that did not exceed six days. There is no reason to believe that there are fundamental differences in the mechanism that silences *Oct4* between the two differentiation protocols. A slowing down of the silencing cascade is a simpler explanation. It is therefore possible that the two opposing results are actually part of one whole story; the slow differentiation protocol points to a delay in *Oct4* downregulation in the absence of DNMT3s and the fast RA-induced differentiation shows that downregulation of *Oct4* is not terminally impeded. In other words DNMT3s and DNA methylation could be acting to enhance the suppressive effects of other factors on *Oct4*.

In the present study, evidence that shows delayed downregulation of *Oct4* comes from the observations in G9a -/- ES cells. EB3-stage embryoid bodies of these cells have significantly higher levels of *Oct4* than their wild-type counterparts

(Figure 3-15) although the final gene shut-down is not affected. Feldman *et al.* (2006) have suggested that G9a acts on the *Oct4* promoter by recruiting DNMT3s in order to permanently silence the downregulated gene. However, the present results do not support this. G9a -/- ES cells could initiate methylation at the proximal enhancer, exactly like the wild type cells, but could not establish the final methylation pattern (Figure 3-13). According to the present study, G9a has a role in accelerating the gene's downregulation, probably by stabilising or enhancing the DNMT3 interaction with the locus.

The conclusions of Feldman *et al.* (2006) were drawn from the fact that differentiated ES cells lacking DNMT3a and b or G9a can re-revert to the pluripotent state –as assessed by *Oct4* expression–, whereas their wild-type counterparts can not. This experiment however does not control for the possibility that the assay could just be selecting for non-differentiated cells that are still expressing *Oct4* at the time of switching to the LIF-supplemented medium. The present analysis suggests that such a resistant population of ES cells that remains undifferentiated under differentiation conditions exists, even in wild-type cells (*Rex1* expression, Figure 3-5 and Figure 3-12, residual *Oct4* expression, Figure 3-6 and Figure 3-15). According to this explanation and in combination with the hypothesis of delayed downregulation stated above, DNMT3a/b -/- and G9a -/- ES cells would have delayed *Oct4* shutdown. At the time of switching medium, more cells would be still expressing *Oct4* in comparison with the wild-type cells. When replated under conditions that support pluripotency, there would be more cells in the KO cell lines that had not yet silenced the gene than in the wild-type cell line and these cells would thrive and give colonies. In other words, the cells were never differentiated and always expressed *Oct4*. One way to test this scenario would be to sort the differentiated ES cells according to an ES-specific surface antigen, such as SSEA-1, before replating. More directly, the number of cells that express *Oct4* could be determined in a subpopulation of the differentiated cells at the time of switching and this number could be later used to normalise the number of *Oct4*-expressing colonies obtained.

Reconstruction of the protein-protein interactions during *Oct4* repression

There are two suppressive mechanisms that are associated with DNMT3s, DNA methylation and association with histone deacetylases (HDACs). Interaction

with HDACs has been shown to be the only means of regulation of PC12 differentiation by DNMT3b (Bai *et al.* 2005). The DNMT3a/b -/- cell line used in this and the other experiments is a catalytic knock-out of the DNA methylatransferase activity in both enzymes (Okano *et al.* 1999). Since the aminoterminal cysteine-rich (Fuks *et al.* 2001) and ATRX-like (Bachman *et al.* 2001) domains of DNMT3s are interacting with the HDACs independently of the catalytic carboxyterminal domain of the protein, there is no reason to assume that the HDAC association in this cell line is in anyway affected. It is therefore likely that the HDAC interactions of DNMT3s are not impaired in the DNMT3a/b -/- ES cells and HDACs can be recruited at the *Oct4* promoter through their interaction with DNMT3s even in the absence of *de novo* methylation.

Gu *et al.* (2006) have shown that MBD3 is recruited to *Oct4* by GCNF regardless of methylation, while MBD2 is also recruited in a methylation-dependent manner later. This methylation-dependent interaction with MBD2 could be the key to the explanation of a delayed downregulation of *Oct4* in the absence of DNMT3s. In the same study, Gu *et al.* (2006) have shown that MBD3 is the main repressor of *Oct4*. Along this line, Kaji *et al.* (2007) have shown that MBD3 -/- embryos have ectopic expression of *Oct4* at the postimplantation stage. MBD2 and 3 are both transcriptional repressors. They are both known to associate with the Mi-2/NuRD complex and the HDACs in it (Zhang *et al.* 1999). At *Oct4*, *MBD3* is recruited first to the locus, and its accompanying NuRD histone remodelling complex could be sufficient for a slow but successful downregulation, while *de novo* methylation and MBD2-related repression could be intensifying the silencing, fine-tuning the initial effect.

Experiments to test this suggested line of interactions that could lead to *Oct4* repression could include sequential immunoprecipitation of MBD3 or MBD2 and components of the NuRD complex in DNMT3a/b -/- differentiating ES cells. If the path of interactions proposed here is correct, then MBD3 but not MBD2 should co-immunoprecipitate with NuRD on the *Oct4* promoter. Moreover, if the association of MBD3 with the NuRD complex is disrupted, silencing would be impossible. Directed mutagenesis of the MBD3 region that is responsible for MBD3 association with the NuRD complex would show if this is true.

"Accelerated repression" of *Oct4*

The discussion above puts forward a model of an "accelerated repression" for *Oct4* (Figure 3-18). According to this model, environmental factors prime the gene for silencing through recruitment of GCNF that displaces the activator SF-1 or LRH-1. GCNF in turn, recruits MBD3 in order to silence the gene. MBD3 is recruited in the context of the NuRD complex (Zhang *et al.* 1999) that has histone deacetylase activity. This recruitment is enough for the downregulation of the gene but at a slow rate imposed by the heterochromatinisation of the locus. HDAC1, which is part of the NuRD complex, is also a partner of DNMT3a (Fuks *et al.* 2001) and could be responsible for its recruitment at the locus. The newly established DNA methylation of the exposed regions stabilises MBD2 localisation to the locus. This model does not make any prediction as to whether MBD2 replaces MBD3 in the NuRD complex or the two complexes coexist, at least transiently, at the promoter of the gene. Recruitment of the MBD2 suppressor intensifies and accelerates the initial repression. HDAC activity on the other hand recruits G9a which methylates the deacetylated H3. G9a can interact with DNMT3a, recruiting more methyltransferase activity to the region. More methylated DNA allows for stronger MBD2 binding and so on. In other words, the downregulation of *Oct4* is happening in waves of repressor recruitment, each wave enforcing the effects of the previous by either recruiting more repressors and/or recruiting more of the same. Whether these factors need to be constantly bound to the gene or the chromatin modification changes are enough to sustain permanent silencing is not known. This model assumes that RNA polymerase II is inhibited by heterochromatinisation and inaccessibility of the locus.

There are a few testable predictions of the "accelerated repression" model. The role of HDACs in accelerating downregulation through their participation in many repressive complexes is central in this model. If the domain of MBD2 that is responsible for HDAC interactions is disrupted, then the downregulation should be slow. Another line of experiments could be based on the prediction that a similar delay in downregulation of the gene should be observed in the absence of either DNMT3s, or MBD2. By repeating the differentiation program in MBD2-/- and DNMT3a/b -/- ES cells, but with much more time course points, *e.g.* every 12 hours, and analysing the course of *Oct4* downregulation, it should be possible to test if this

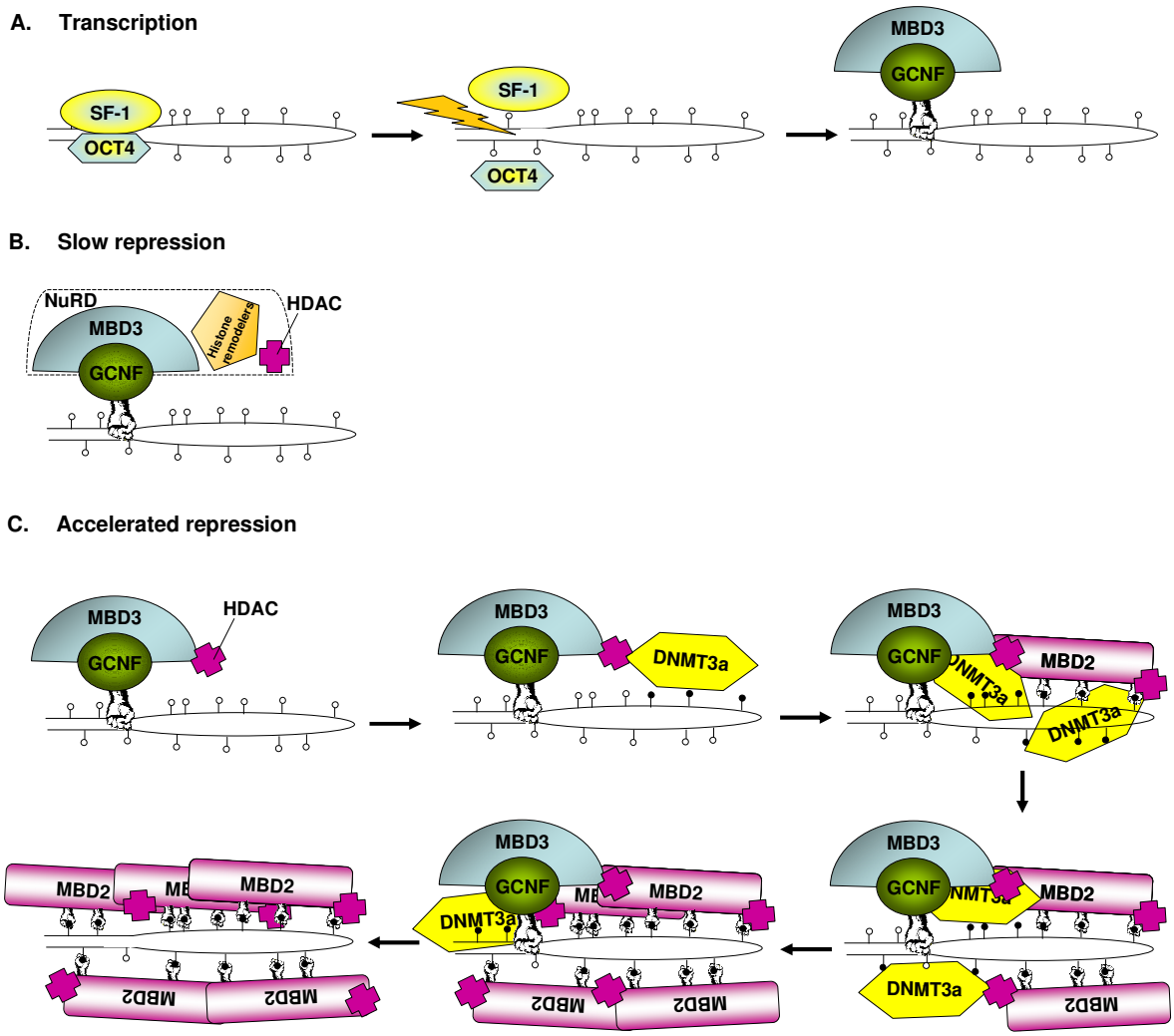

Figure 3-18. Model of accelerated repression of Oct4. (A) In the presence transcription factors the gene is active. Once there is a stimulus to initiate repression, these factors are replaced by GCNF that is loaded with the repressor MBD3. In the absence of other interactions, transcription continues. (B) The repression starts because MBD3 has been recruited with the NuRD complex that contains HDACs and chromatin remodelers that make the chromatin environment hostile to transcription. This downregulation is slow but enough to lead to the complete repression on the gene. (C) The histone deacetylases of the NuRD complex can form a complex with DNMT3a, recruiting it to the locus. DNA methylation of unprotected regions stabilises MBD2 binding to the locus. MBD2 can also interact with histone deacetylases that in turn attract DNMT3a and the cascade escalates causing quick downregulation of the gene. Although the model has been depicted with the distal enhancer and the promoter forming a loop, this is not important for its validity. Empty circles are CpGs and filled circles are methyl-CpGs. Finally, all the protein-protein interactions are considered reversible, following the rules of equilibrium dynamics.

is indeed the case. It would be interesting to repeat the above experiment in a cell line that is deficient of only DNMT3a and see if DNMT3b can substitute for its action. Finally, the model predicts that there is a stoichiometric increase in HDACs and MBD2 as the gene downregulation proceeds. Carefully controlled chromatin immunoprecipitation experiments should be able to test for such an increase.

4. Investigation of methylation in mouse CpG islands

4.1. CpG islands in the mammalian genome

4.1.1. Discovery and definition of CpG islands

In the early eighties it was known that vertebrate genomes have high CpG methylation levels and that the mammalian genome is generally depleted of CpG dinucleotides, most probably due to the hypermutability of methylcytosines (Bird 1980; Duncan and Miller 1980). Naveh-Many and Cedar (1982) and Cooper *et al.* (1983) independently showed that despite the overall high methylation, there is a fraction in the vertebrate genome that is virtually methylation-free. More detailed analysis of some of these methylation-free genomic fragments (Bird *et al.* 1985) demonstrated that they are also GC-rich in comparison to the rest of the genome and have increased observed versus expected frequency of CpGs (o/e). Regions of high o/e in the genome had been already recognised as being preferentially found in the 5' regions of genes (McClelland and Ivarie 1982). Bird (1986) finally identified the 5'-associated CpG-rich regions with the unmethylated "CpG-rich islands". The first formal definition of CpG islands came from Gardiner-Garden and Frommer (1987) that defined them as a "*200-bp (or longer) stretch of DNA with a GC content of 50% and an o/e in excess of 0.6*".

4.1.2. Distribution of CpG islands in the genome

The genomes of mouse and human are reported to contain about 15,500 and 27,000 CpG islands respectively, distributed on the chromosomes with an average density of 5-10 CpG islands/ Mb (Mouse Genome Sequencing Consortium 2002). The only exception to this rule is human chromosome 19, which is extremely CpG island-rich and has approximately 45 CpG islands/ Mb. The distribution of CpG islands on the chromosomes is not uniform and they appear to colocalise with what are known as G bands (Craig and Bickmore 1994). Since the human and mouse genomes are estimated to contain around 22,000 genes, it would appear —at least for human— that CpG islands are in excess in comparison to genes. Moreover, it has long been known that not all genes are associated with CpG islands. In more detail, it is estimated that around 50% of the mouse genes and 56% of the human genes

associate with a CpG island (Antequera and Bird 1993). This leaves an estimate of 34% and 55% of all CpG islands in mouse and human respectively to be intergenic. The high number of intergenic CpG islands in humans is enough to explain the difference between human and mouse regarding the total number of CpG islands.

CpG islands in the 5' region of human genes are often flanked by Alu elements that also have a GC-rich DNA sequence (Kang *et al.* 2006). By applying the CpG island criteria of Gardiner-Garden and Frommer (1987) on human chromosomes 21 and 22, Takai and Jones (2002) found that about half of the identified CpG islands had recognisable Alu motifs. This proportion was four times smaller when more stringent CpG island selection criteria (500 nt, 55% GC and 0.65 o/e) were applied and the new CpG island criteria were suggested as more accurate.

According to data on human chromosomes 21 and 22, 17.4% of the CpG islands are located in the 5' region of a gene (Takai and Jones 2002). CpG islands that lie at the 5' of genes can begin outside of the transcribed region, or within the first exon, but most of them extend well into the gene (Gardiner-Garden and Frommer 1987). CpG islands in the 5' are not depleted for TATA boxes and, as expected from their sequence composition and they often contain GC-boxes (Gardiner-Garden and Frommer 1987).

Using the genome sequence of mouse and human, it has been shown that CpG islands are preferentially, but not exclusively, present in the promoters of housekeeping genes (Yamashita *et al.* 2005; Saxonov *et al.* 2006). This has been associated with the negative effect that DNA methylation has on transcription and the need for housekeeping genes to be constitutively expressed. Robinson *et al.* (2004) explored the possible functions of CpG islands that appear to be present in tissue-specific genes. They compared CpG island-associated and non-CpG island-associated genes with respect to the gene ontology (GO) terms accompanying them and their relative abundance in embryonic expressed sequence tag (EST) libraries. They made the interesting observation that many of the tissue-specific genes that are associated with CpG islands are important during development. Finally, earlier work on a limited number of genes with CpG island promoters and various expression patterns indicated that the CpG islands of tissue-specific genes tend to be at the lower

end of the o/e spectrum for CpG islands (Edwards 1990). The same result was later confirmed for genes that have a role in development (Ponger *et al.* 2001).

13% of CpG islands (Takai and Jones 2002) are associated with exons, and a remaining small fraction appears to be at the 3' of genes (Gardiner-Garden and Frommer 1987). Because base composition in the exons is important for carrying the genetic information, it has been hypothesised that CpG-rich exonic sequences may be selected because of their coding potential for arginine (CGN and AG(A/G)). This scenario relies on the accepted fact that selection acts against mutations that would have deleterious effects on the encoded proteins. In the case of arginine, selection would act against mutated CpGs, which in turn would lead to the maintenance of the o/e of the region. Gardiner-Garden and Frommer (1987) tested this hypothesis in a big variety of exon-associated CpG islands in a variety of vertebrates. They calculated the o/e of these regions with and without the arginine codons and they found that the high o/e content of them could not be attributed solely to codon usage for most of the genes.

A more generalised theory trying to explain the presence of CpG islands in the gene body is the isochore theory. According to this theory, housekeeping genes are embedded in GC-rich regions and this is reflected in the base composition of their wobble third codon position that is not under strict selection. Analysis of the base composition in the gene body of housekeeping and tissue-specific genes has been contradictory (Pesole *et al.* 1999; Goncalves *et al.* 2000; Duret and Galtier 2000; Ponger *et al.* 2001). Other analyses suggest that the isochore theory might be true for human but not mouse (Vinogradov 2003).

There is very little research on the 3' CpG islands. Shabalina *et al.* (2003) have shown that GC-rich 3' untranslated regions (UTR) tend to be conserved between human and mouse and they suggested that the base composition might be important for their function during transcription. However, they did not specifically examine CpG islands in these positions. In another report, comparative analysis of the non-coding regions of homologous mouse and human genes showed high conservation at their 3'UTRs (Jareborg *et al.* 1999). A CpG island was associated with the 3' UTR of 10% of the mouse and 35% of the human genes. Much more

experimentation and analysis is required in order to understand the role of 3' CpG islands.

4.1.3. Methylation status of CpG islands

Because of the history of their discovery, CpG islands are typically thought of as devoid of methylation. This assumption was evident in the construction of the first comprehensive CpG island library by selecting genomic sequences mainly on the basis of them being free of methylation (Cross *et al.* 1994). Examples such as imprinted genes, X-linked CpG islands and aberrantly methylated cancer cells show however that there is no evidence that CpG islands are intrinsically "unmethylatable". Indeed, there is a long line of evidence for the presence of methylated CpG islands in the mammalian genome.

Tissue-specific CpG island methylation patterns

The first genome-wide approaches for the investigation of CpG island methylation were based on a methylation-sensitive restriction digestion protocol, namely restriction landmark genome scanning (RLGS). RLGS experiments showed that, at least in mouse, there is a dynamic methylation profile characteristic of the cell type. Comparison of the methylation profile in a variety of mouse tissues showed that approximately 5.6% of the CpG islands show differential, tissue-specific methylation patterns (Song *et al.* 2005). Importantly, although there appeared to be an overall methylation signature for each tissue, the methylation status of each of these loci could be shared among more than one tissues.

The fact that all the different tissues arise from the differentiation of a single fertilised oocyte, makes it obvious that these different methylation patterns have to be brought about during development. RLGS experiments that were testing this programmed CpG island methylation hypothesis showed that, indeed, there is a characteristic methylation profile that changes between ES cells, trophoblast cells, embryonic germ cells, embryoid bodies, the foetus and somatic tissues (Shiota *et al.* 2002; Kremenskoy *et al.* 2003). Perhaps the most important discovery of these analyses is that, although the total number of methylated loci increases with differentiation as expected, there is not a general trend for increasing methylation in each locus, but the locus can become methylated or lose methylation to give rise to

the tissue-specific methylation pattern. Finally, it appears that the earlier stages of differentiation are the time point in which most of the tissue-specific methylation reprogramming events happen. This is supported by the fact that during differentiation, the differentially methylated loci were approximately 16% of the total (Shiota *et al.* 2002) as opposed to only 5.6% when adult tissues were compared (Song *et al.* 2005). Additionally, RLGS comparison of the final stages of brain development (Kawai *et al.* 1993) showed that only 1.7% of the CpG islands were differentially methylated in this later developmental stage to give rise to the adult brain.

Comparison of the methylation status of CpG islands in various healthy human tissues has showed that, like in mice, there is a tissue-specific pattern of CpG island methylation and some of these differentially methylated CpG islands are shared between different tissues (Eckhardt *et al.* 2006 and R. Illingworth, personal communication). Interestingly, in most of the studies, CpG island methylation seems to be restricted to those regions that are in the lower spectrum of o/e values, something that probably is due to increased mutation rates of methylated CpGs (Fang *et al.* 2006; Bock *et al.* 2006; Weber *et al.* 2007). Additionally, Weber *et al.* (2007) have shown that methylation of CpG island promoters correlates well with lack of transcription, especially in the CpG islands with the highest o/e values. However, in this study, methylation was also detected in active promoters with low o/e. Detailed methylation profiling of human chromosomes 6, 20 and 21 has shown that CpG islands located at gene promoters are more resistant to methylation (Eckhardt *et al.* 2006) ; 12.1% of the CpG islands that were associated with the 5' of genes were methylated, in contrast to 23.4% of non-5' associated CpG islands.

An interesting case of tissue-specific CpG island methylation in human is that of the *ψSLC6A8* pseudogene. The 5' CpG island of the normal *SLC6A8* gene in this case is free of methylation, but that of the pseudogene is methylated in all the studied tissues except from testes (Grunau *et al.* 2000). This is possibly a mechanism employed by the cell to restrict transcription of the non-functional gene that can probably respond to the same transcription factors as the normal gene.

CpG island methylation patterns in sperm

A specific case of tissue-specific CpG island methylation is that of sperm. The reason for this separation from the other tissues is the hereditary potential of these cells. It has been postulated that sperm (as well as oocytes, which are much more difficult to study) should have lower global levels of DNA methylation because any spontaneous deleterious mutation of methylcytosines could be transmitted to the offspring. Moreover, many consider germ cells as being totipotent (although another school of thought supports they are highly differentiated) and their expression potential therefore needs to be free of the transcription constraints imposed by methylation. Perhaps surprisingly, human sperm appears to have significant CpG island methylation, although, in overall, it is lower than that of somatic tissues (Weber *et al.* 2007). As in somatic tissues, the CpG islands in the higher end of o/e values have the least representation in the methylated fraction. Importantly, germline-specific genes were depleted from the methylated fraction of CpG islands. However, the majority of methylation in testes is found in non-CpG island and repetitive sequences (Oakes *et al.* 2007b).

CpG island methylation and ageing

It has been postulated for some time that CpG island methylation increases with age, something that has been related to the aberrant methylation phenotype of cancer cells (Ting *et al.* 2006). However, the experimental evidence to support this hypothesis is fragmental. Of the recent experimental data, a study on monozygotic twins has shown that although the siblings were epigenetically indistinguishable in the beginning of their life, they accumulated DNA methylation as they aged (Fraga *et al.* 2005). This study however did not examine specifically CpG islands. Another study showed that the CpG islands associated with the *CSX* and *SOX10* human genes gain methylation with age (Chu *et al.* 2007). The authors could explain this tendency with a mathematical model in which random methylation errors accumulated during repeated mitotic cell divisions. On the opposite side, the study from Eckhardt *et al.* (2006) did not reveal any significant differences in the global DNA methylation levels when different age groups were compared. The effect that ageing could have

on the methylation landscape of the genome is a very interesting concept with many applications in medicine that should be more systematically investigated.

4.1.4. Dynamic evolution of CpG islands

What is causing the existence of CpG islands in vertebrate genomes? The answer to this question is not known but speculations can be made. DNA methylation is associated with cytosine loss through mutation (Bird 1980; Duncan and Miller 1980). This relationship between cytosine methylation and cytosine depletion is evident in the genomes of invertebrates that do not have DNA methylation and have a high frequency of the CpG dinucleotide, as well as in the observation that CpG islands of lower o/e are more often found methylated than those of high o/e. According to this, it would be perhaps more appropriate to view CpG islands not as regions with increased CpG frequency, but rather as regions in the CpG-depleted genome where the CpG frequency approaches the theoretically expected. Following this line of thought, the theoretically expected base composition of the genome should have been present in an ancestral organism, before the appearance of DNA methylation in the genome in its present form. If this is true, then the presence of CpG islands could be linked to an ancestral tendency of these regions to avoid methylation which is evident until today.

By studying the presence of methylation in the different categories of eukaryotes (Figure 4-1) one can see that DNA methylation is present in organisms as diverse as plants, protists, and chordates. It would appear that being a eukaryote goes together with having DNA methylation and its absence from the genome of certain eukaryotes is acquired at a later stage, independently. If this is the case, then some traces of CpG depletion (which is the biological footprint of CpG methylation) should be present today in these organisms. Indeed, the genome of *D. melanogaster* that does not have DNA methylation has an average o/e of approximately 0.7 and not 1, as it would be expected if there were no bias (Jabbari and Bernardi 2004). Interestingly, both the genomes of *D. melanogaster* as well as that of *S. cerevisiae* show CpG peaks around the promoters of genes which resemble the CpG islands of vertebrates (Shimizu *et al.* 1997). It could be hypothesised that the biological force that caused the formation of CpG islands in vertebrates was in action before these

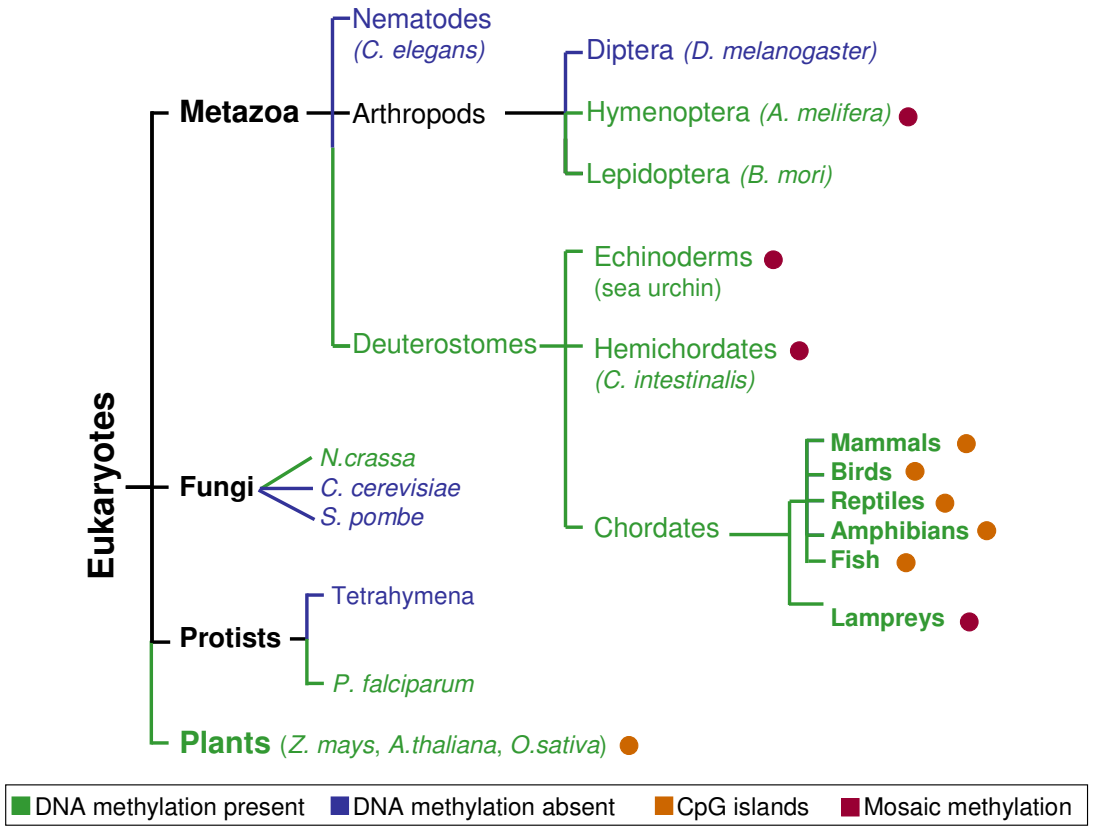

Figure 4-1. Schematic phylogenetic relationships of various eukaryotes and the occurrence of CpG islands. The key to the colour-codes is shown below the figure. A mosaic methylation pattern is smaller regions of CpG depleted DNA embedded in a CpG-rich genome. In these cases, the CpG depleted regions colocalise with dense methylation. Representative organisms in each category are shown in parentheses. The list of organisms is by no means extensive. There is no information about the CpG distribution in the genome of protists. Neurospora has DNA methylation in specific sequences that tend to be CpG-depleted. (Pollack *et al.* 1991; Tweedie *et al.* 1997; Shimizu *et al.* 1997; Regev *et al.* 1998; Simmen *et al.* 1999; Varriale and Bernardi 2006a; 2006b and "The tree of life web project")

organisms diverged from the common ancestor and was frozen as soon as methylation vanished from these genomes.

What could be the benefit of resisting methylation in an otherwise globally methylated genome? As discussed thoroughly in the introduction, DNA methylation seems to be refractory to transcription. Perhaps, in order to ensure transcription of certain genes, the cell developed a mechanism to prevent methylation in these promoters. This is supported by the fact that the majority of housekeeping genes is associated with a CpG island in the promoter region (Yamashita *et al.* 2005; Saxonov

et al. 2006), and that a big proportion of CpG islands is located at the 5' of genes (Gardiner-Garden and Frommer 1987; Takai and Jones 2002). Moreover, analysis of promoter-associated CpG islands in mouse, human and dog has shown that orthologous genes in these organisms tend to maintain their association with CpG islands (Jiang *et al.* 2007). Given all the evidence, transcription regulation could be a reason that certain genomic regions remained methylation-free and gave rise to today's CpG islands.

The above hypothesis however, does not provide any explanation about the presence of CpG islands in non-regulatory and intergenic regions. It has been proposed that CpG islands serve as origins of replication (ORIs) (Delgado *et al.* 1998; Antequera and Bird 1999). This theory has been put forward in order to explain the absence of methylation in CpG islands and is based on observations on CpG islands that are associated with 5' of genes. Gomez and Antequera (1999) have shown that in *S. pombe* replication firing does not depend on transcription and ORIs can be found at intergenic regions. There is no reason, to my knowledge, that the model of ORI-associated CpG islands can not be expanded to include non-promoter CpG islands.

As illustrated in Figure 4-1, not all the organisms that have DNA methylation have the same pattern of CpG depletion. Organisms like *C. intestinalis* have mosaic methylation of CpG-depleted and methylated "islands" in an otherwise CpG-rich and unmethylated genome (Simmen *et al.* 1999). This raises the possibility that a different mechanism is in action for the formation of the CpG pattern in the genomes of invertebrates. Based on the observation that DNA methylation and CpG depletion is targeted to the gene body in *C. intestinalis,* it has been suggested that methylation in the gene body could be preventing aberrant transcription from cryptic promoters (Suzuki *et al.* 2007). This could be an alternative mechanism that regulated the CpG distribution that is characteristic of all the other deuterostomes except from mammals, birds, reptiles, amphibians and fish. *A. thaliana* seems to both have CpG islands and show transcription-related gene body methylation (Zilberman *et al.* 2007). There is no reason to exclude the possibility that these two putative mechanisms of formation of CpG-rich regions acted in parallel for the formation of the CpG patterns observed in today's organisms.

4.1.5. Remaining questions

The existence of CpG islands and their methylation characteristics in mammals and other organisms pose a series of biologically important questions. First of all, what has been the force that led to their formation? Related is the question of why some organisms have evolved to harbour DNA methylation and others not? What determines whether a CpG island will be methylated and what is the cellular mechanism that ensures it will remain methylation-free? What is the role of CpG islands that are not located at promoters? Is CpG island methylation determining the developmental fate of a cell?

The goal of this study was to identify for the first time the CpG island methylation patterns in the mouse brain and to investigate to which extent these patterns are being established during development. In more detail, the CpG islands in the mouse genome were computationally identified and their sequences were used for the preparation of a mouse CpG island tilling microarray. Hybridisation to the CpG island microarray was then used for the identification of the CpG islands that are methylated in brain, which had been isolated using an affinity purification technology. Additionally, the same methodology was used for the identification of CpG islands that become *de novo* methylated during *in vitro* differentiation of ES cells and the results were compared to the CpG island methylation in brain. The implications of the results on the establishment of the CpG island methylation during development in mouse are discussed.

4.2. Identification of the CpG islands in the mouse genome

For the purposes of constructing a CpG island microarray for this study, the CpG islands of mouse were computationally identified. This was done by using the more relaxed 0.6 o/e and 50% GC thresholds as the CpG islands of mouse are less CpG-rich than the human ones (Matsuo *et al.* 1993). To avoid however contamination with low complexity sequences, the genome sequence was first

masked and the repetitive and low complexity sequences were discarded before filtering for CpG islands. Furthermore, to improve the specificity of the screen, the minimum length requirement was 500 nucleotides. This method returned sequences that *a priori* do not include repetitive or low complexity elements and have the typical characteristics of CpG islands (Figure 4-2).

The original purpose for the identification of the CpG islands of mouse had been the construction of a microarray that was intended to be used in conjunction with affinity purified CpG islands. As it will be explained more thoroughly later (section 4.3.3), the fragmentation of the genome for the purification of the CpG islands had been performed by means of Mse I restriction digestion. This dictated that the algorithm for the identification of the CpG islands was applied on an *in silico* Mse I-digested genome. In order to reassemble possibly fragmented CpG islands, neighbouring Mse I fragments that passed the CpG island selection criteria were grouped together if they were separated by less than 200 bp and/or they spanned a region equal or less than 2,500 nucleotides. After this, the total number of CpG islands present on the array was calculated to be 20,755.

4.2.1. Distribution of CpG islands in the mouse genome

24.7% of the CpG islands are intergenic. The distribution of the CpG islands on the chromosomes shows a strong positive correlation with their gene content (Pearson's correlation: 0.9, $p=1.85\ 10^{-8}$, Figure 4-3). The only exception is chromosome X that shows fewer CpG islands than expected by its gene content. Of the gene-associated CpG islands, 67.34% (50.71% of all the CpG islands) are at the 5', 6.3% (4.75% of all the CpG islands) are at the 3' and 30.88% (23.25% of all the CpG islands) are intragenic. A simplified diagram that shows the distribution of CpG islands relative to the transcription start site is shown in Figure 4-4.

The number of genes that are associated with the CpG islands (in any position relatively to the transcription start site) is 62.87% of the total (NCBI build 34). If only the 5' CpG islands are taken into account, then the percentage of genes that are associated with a CpG island is 51.72%. This number is very close to the one described before (49.6%, Antequera and Bird 1993).

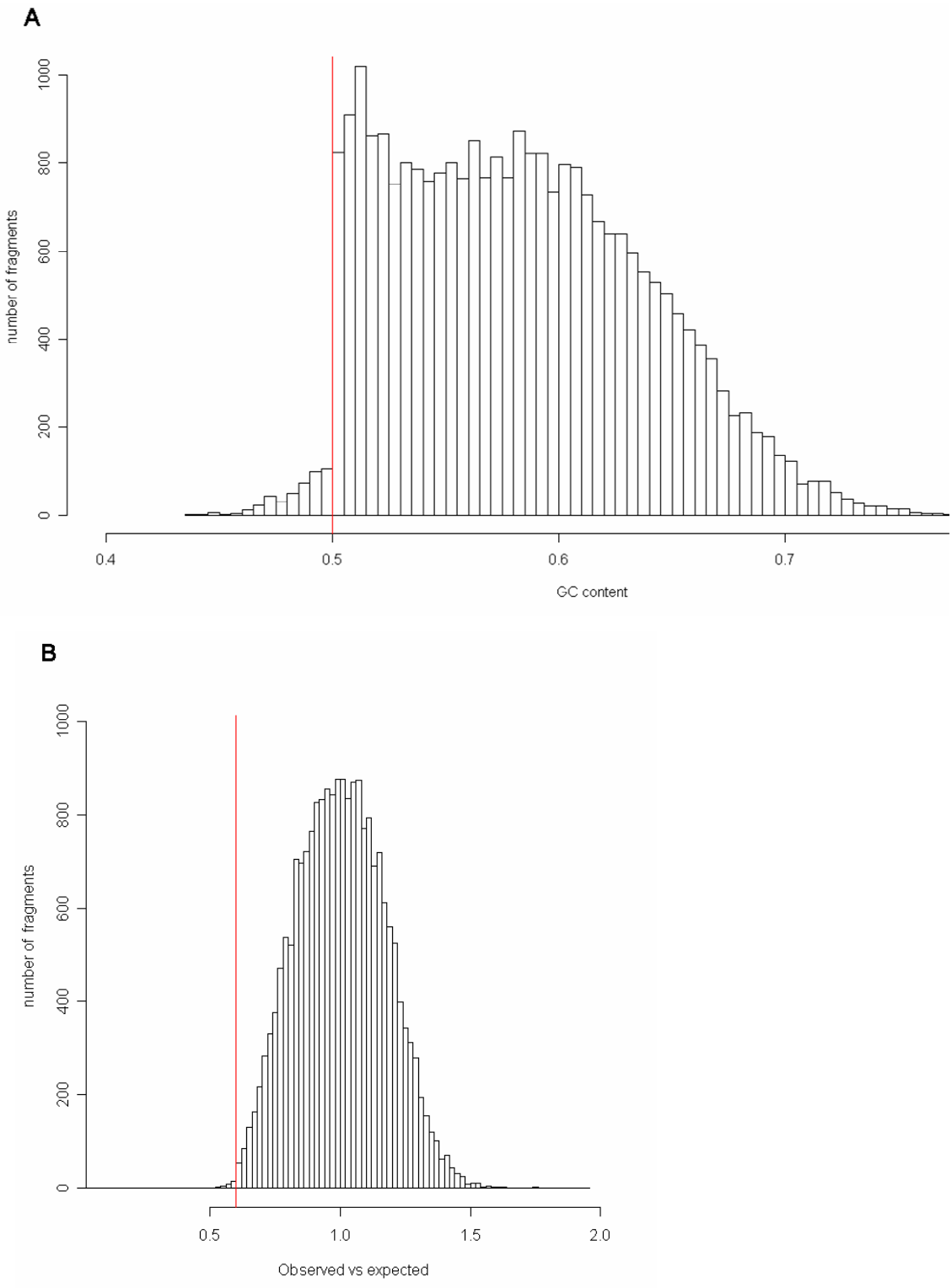

Figure 4-2. Histograms of GC-richness and o/e of the Mse I fragments on the microarray. (A) Distribution of the Mse I fragments according to their GC content. The vertical red line shows the 0.5 threshold of the CpG island definition. (B) Distribution of the Mse I fragments according to their o/e value. The o/e was calculated along 200 nucleotide windows. The vertical red line shows the 0.6 threshold of the CpG island definition.

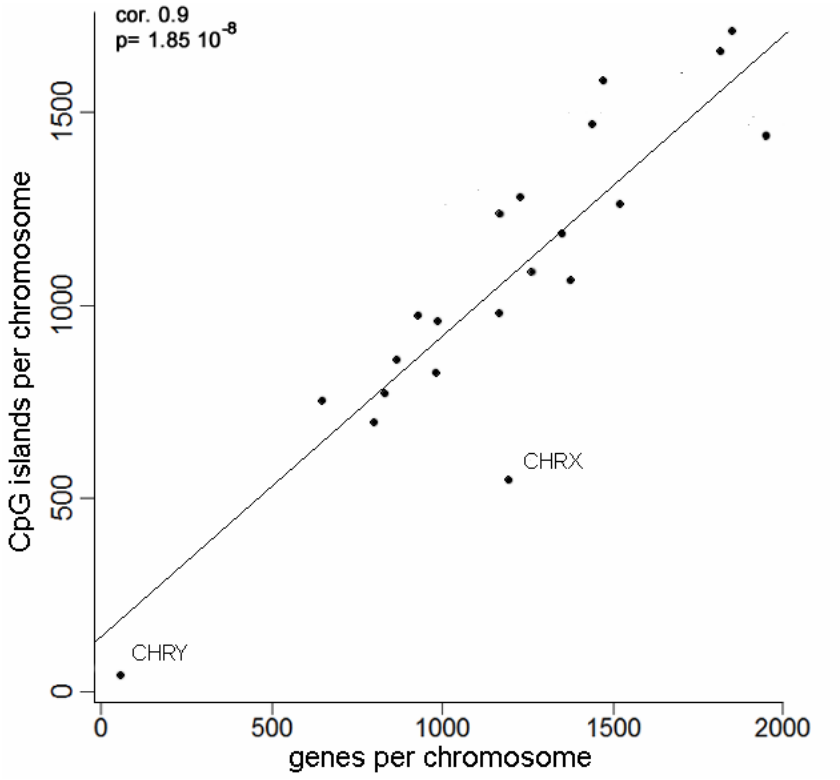

Figure 4-3. Scatter plot of the number of CpG islands against the number of genes in each chromosome. The regression line is also shown. The Pearson's correlation and the confidence value for the two variables being positively correlated is shown at the top left. Chromosomes X and Y are marked.

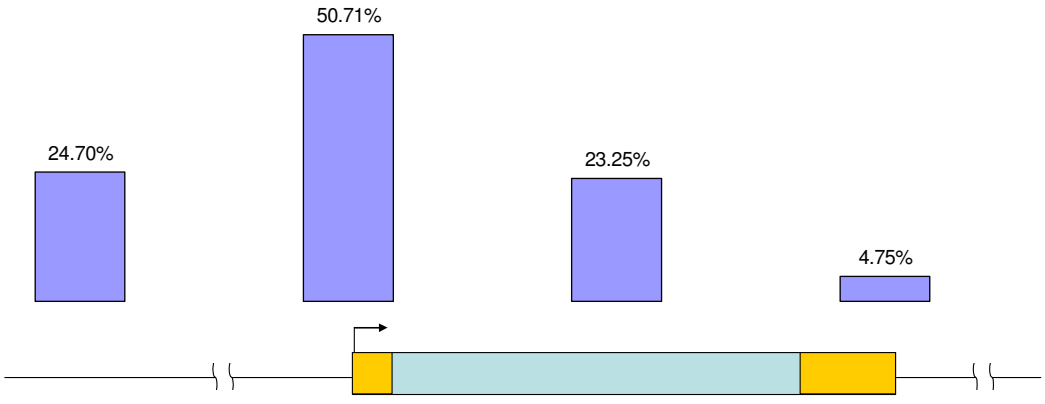

Figure 4-4. Simplified diagram of the distribution of CpG islands relative to the transcription start site. The vertical bars show the frequency of CpG islands in intergenic regions, 5', intragenic regions and 3' (left to right). A cartoon of the respective regions is shown at the bottom. The thin line represents intergenic regions and the box genes. The arrow shows the transcription start site, yellow boxes show 5' and 3' untranslated regions and the pale blue box the translated region (introns and exons).

The information on the CpG island distribution in the mouse genome and their association with genes is shown in Table 4-1.

An interesting observation can be made when the number of CpG islands per gene region is calculated. In total, there appear to be on average 1.061 CpG islands per gene which are distributed as follows: there are 0.82 CpG islands per 5', approximately one CpG island per intragenic region and 1.3 CpG islands per 3'. The overlap of one CpG island with the 5' of more than one genes was further investigated. 10% of all the 5'-associated CpG islands were shared between two genes, and in almost all of these cases, the genes were of opposite orientation. Such gene arrangements are not rare in mammals and raise possibilities for a common regulatory mechanism.

Table 4-1. Distribution of CpG islands in the mouse genome.

	Number of CpG islands	**CpG islands (%)**	**Gene-associated CpG islands[1] (%)**	**Number of CpG island-associated genes**	**CpG island-associated genes[2] (%)**	**CpG island-associated genes[3] (%)**
Intergenic	5126	24.70	NA[4]	NA[4]	NA[4]	NA[4]
5'	10524	50.71	67.34	12798	82.42	51.73
Intragenic	4826	23.25	30.88	4899	31.55	19.8
3'	985	4.75	6.30	785	5.06	3.17
Total	*20755*[5]	*100*[5]	*100*[5]	*15528*[5]	*100*[5]	*62.87*

[1] The number of CpG islands in each gene region divided with the total number of gene-associated CpG islands (5', Intragenic and 3').

[2] The number of CpG island-associated genes in each category (5', Intragenic and 3') as a fraction of the total number of CpG island-associated genes calculated by substracting the intergenic CpG islands from the total.

[3] The number of CpG island-associated genes in each category (5', Intragenic and 3') as a fraction of the total number of genes predicted in NCBI Build 34 (*i.e.* 24741 genes*)*.

[4] Not applicable.

[5] This number is smaller than the sum of the rows above because of the cases of CpG islands mapping in more than one gene regions.

4.2.2. Functional annotation of the genes that are associated with CpG islands in mouse

The genes that were associated with the CpG islands were compared against the NCBI gene library using the PANTHER interface (http://www.pantherdb.org). The probability that a specific class of genes is enriched in the library was corrected for multiple testing with the Bonferroni algorithm. The gene classes identified as being preferentially associated with CpG islands can be seen in Table 4-2 and include development, metabolism and protein synthesis, cell cycle, signalling, transport, cell structure as well as some genes that are important for neuronal activity. The terms "cell cycle", "metabolism", "developmental processes", "protein transport", "protein modification", "cell communication", "transport", "death"

Table 4-2. Gene ontology categories that are significantly enriched in CpG island-associated genes.

	GO term	No. of genes in NCBI	No. of genes with CpGi	Expected	P value
Developmental	Developmental processes	2385	1040	664.79	$6.25\ 10^{-44}$
	Neurogenesis	636	349	177.28	$1.13\ 10^{-28}$
	Ectoderm development	729	375	203.2	$6.90\ 10^{-26}$
	Mesoderm development	576	278	160.55	$1.84\ 10^{-15}$
	Cell proliferation and differentiation	1004	420	279.85	$3.36\ 10^{-14}$
	Skeletal development	131	82	36.51	$1.20\ 10^{-8}$
	Embryogenesis	167	90	46.55	$1.39\ 10^{-6}$
	Anterior/posterior patterning	68	41	18.95	$1.08\ 10^{-3}$
	Heart development	52	33	14.49	$4.07\ 10^{-3}$
	Other developmental	135	63	37.63	$1.37\ 10^{-2}$

Table 4-2. (Continued)

	GO term	No. of genes in NCBI	No. of genes with CpGi	Expected	P value
Metabolism	Nucleoside, nucleotide and nucleic acid metabolism	3851	1432	1073.42	$4.65\ 10^{-28}$
	Protein metabolism and modification	3819	1225	1064.5	$4.06\ 10^{-6}$
	Carbohydrate metabolism	608	240	169.47	$4.45\ 10^{-6}$
	Lipid, fatty acid and steroid metabolism	879	326	245.01	$9.63\ 10^{-6}$
	Phosphate metabolism	118	60	32.89	$4.17\ 10^{-4}$
	Amino acid metabolism	248	104	69.13	$1.59\ 10^{-3}$
	Phospholipid metabolism	148	71	41.25	$2.24\ 10^{-3}$
	DNA metabolism	327	137	91.15	$5.80\ 10^{-4}$
	Proteolysis	1151	394	320.83	$4.42\ 10^{-3}$
	Other metabolism	627	251	174.77	$7.72\ 10^{-7}$
Protein synthesis	Protein phosphorylation	750	335	209.05	$5.26\ 10^{-14}$
	Protein modification	1300	569	362.36	$7.57\ 10^{-23}$
	Protein glycosylation	192	91	53.52	$3.55\ 10^{-4}$
	mRNA transcription	2275	845	634.13	$7.02\ 10^{-15}$
	mRNA transcription regulation	1687	645	470.23	$3.74\ 10^{-13}$
Cell cycle	Cell cycle control	425	203	118.46	$1.05\ 10^{-10}$
	Mitosis	374	152	104.25	$8.70\ 10^{-4}$
	Apoptosis	544	231	151.63	$2.71\ 10^{-8}$
	Induction of apoptosis	161	79	44.88	$3.54\ 10^{-4}$
	Homeostasis	227	91	63.27	$1.81\ 10^{-2}$

Table 4-2. (Continued)

	GO term	No. of genes in NCBI	No. of genes with CpGi	Expected	P value
Signaling	Signal transduction	4469	1627	1245.68	$2.89\ 10^{-28}$
	Other receptor mediated signaling pathway	207	106	57.7	$1.36\ 10^{-6}$
	Cell surface receptor mediated signal transduction	2601	859	725	$3.03\ 10^{-5}$
	Receptor protein tyrosine kinase signaling pathway	218	109	60.76	$2.80\ 10^{-6}$
	Cell adhesion-mediated signaling	449	178	125.15	$8.47\ 10^{-4}$
	Cell communication	1257	525	350.37	$4.68\ 10^{-17}$
	Intracellular signaling cascade	909	420	253.37	$1.88\ 10^{-20}$
	Other intracellular signaling cascade	233	105	64.95	$5.43\ 10^{-4}$
Transport	Transport	1455	547	405.56	$1.26\ 10^{-10}$
	Cation transport	527	204	146.89	$7.85\ 10^{-4}$
	Protein targeting and localization	217	92	60.49	$2.86\ 10^{-3}$
	Intracellular protein traffic	1044	424	291	$1.74\ 10^{-12}$
Neuronal	Other neuronal activity	132	73	36.79	$1.24\ 10^{-5}$
	Synaptic transmission	284	117	79.16	$5.50\ 10^{-3}$
Cell structure	Cell structure and motility	1128	459	314.42	$1.14\ 10^{-13}$
	Cell structure	678	272	188.98	$7.96\ 10^{-7}$
	Cell adhesion	672	272	187.31	$7.79\ 10^{-8}$

and “signal transduction” have also been shown to be preferentially associated with CpG islands in independent studies in mouse and human (Yamashita *et al.* 2005; Saxonov *et al.* 2006).

The developmental and housekeeping functions of the CpG island-associated genes vastly outnumber the rest. An association of CpG islands with housekeeping genes was noticed since early, but it is only recently that a role in development has started to be recognised. The preferential association of CpG islands with genes involved in signalling has not been reported before and is very interesting as this function is largely cell-type dependent. Another highly specialised role of genes that are associated with CpG islands is a role in neuronal function. Further detailed investigation of the genes that consist these functional groups could provide more insight into their significance.

4.3. MBD-affinity purification (MAP) of methylated CpG islands

MAP relies on the affinity of the MBD domain for methyl-CpGs and is based on the protocol of Cross *et al.* (1994). Briefly, total genomic DNA is digested with Mse I (TTAA) that preferentially cuts outside CpG island sequences. Then, methylated CpG islands are purified using an MBD domain affinity purification column. The purified DNA is finally hybridised to a microarray that contains sequences representing all the CpG islands. The hybridisation signal is interpreted in terms of enrichment or depletion of any given CpG island relative to the initial, total DNA.

4.3.1. Murine CpG island microarray

At the time this work was carried out there was no commercially available microarray, specific for mouse CpG islands. The CpG islands in the mouse genome were identified as described in section 4.2 and were flanked by Mse I recognition sites (TTAA). The 26,687 Mse I fragments that passed the CpG island criteria were

used for the construction of the oligonucleotide microarray. Each fragment was represented on the array by an average number of 14.4 50-mer probes that were tiled in 48-nucleotide intervals. The total number of probes on the array was 385,215.

4.3.2. Purification of the MBD protein and packing of the column

The MBD protein was expressed and purified as described in the Materials and Methods. The expected molecular weight of the MBD protein is 10.4 KDa and the process of the purification was monitored with SDS-PAGE (Figure 4-5). The purified protein was analysed with EMSA to confirm that its methyl-DNA binding activity was not compromised (Figure 4-6). Next, the purified, recombinant protein was bound to nickel beads and packed into a column. The efficiency of the binding to the beads was evaluated with SDS-PAGE (Figure 4-7).

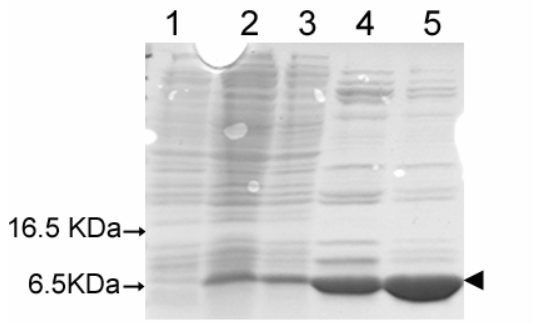

Figure 4-5. SDS-PAGE monitoring of the MBD purification process. A representative example is shown. Lane 1, $1:10^{-5}$ v/v of uninduced bacterial culture; Lane 2, $1:10^{-5}$ v/v of bacterial culture induced for recombinant protein expression for 2 hours with IPTG; Lane 3, $1:10^{-5}$ v/v of bacterial lysate; Lanes 4 and 5, 1:600 v/v of first and second elution of the purified protein. The arrowhead shows the band that corresponds to the MBD protein.

4.3.3. Preparation of the DNA

Genomic DNA was digested with Mse I (TTAA). This was the chosen method for the fragmentation of the genomic DNA since Mse I recognises sequences outside CpG islands more frequently and is less likely to destroy them (Cross *et al.* 1994). The digestion was carried to completion and then the DNA was

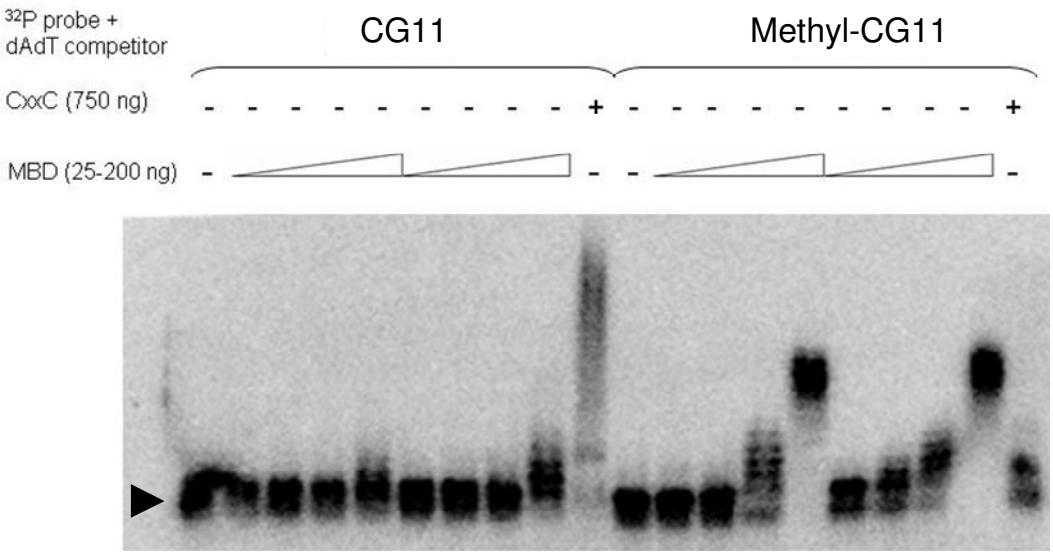

Figure 4-6. EMSA of the purified MBD protein. The methyl-binding activities of two independent protein preparations are shown. Increasing amounts of the MBD protein were incubated with either the unmethylated or the *in vitro* methylated CG11 radiolabelled probe in the presence of 1µg of the non-specific competitor poly-dAdT. Controls with the CxxC protein that binds specifically umethylated CpGs are included. The arrowhead shows the free probe.

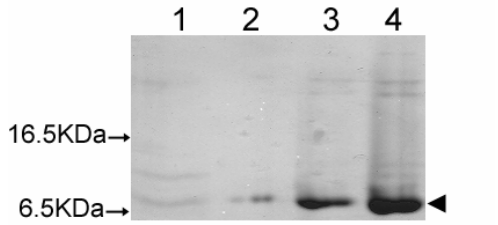

Figure 4-7. SDS-PAGE evaluation of the binding efficiency of the purified MBD protein to Ni-NTA sepharose beads. A representative example is shown. Lane 1, 1:750 v/v of the unbound protein; Lanes 2-4, increasing volumes of protein that is bound to the Ni-NTA sepharose beads. The arrowhead shows the band that corresponds to the MBD protein.

dephosphorylated. The dephosphorylation was tested for completeness by incubating the sample with T4 ligase for one hour and resolving the DNA with agarose electrophoresis (Figure 4-8, A). Next, the adaptors were ligated to the DNA and the efficiency of the ligation was tested by PCR (Figure 4-8, B). The prepared DNA was stored at -20^{o} C.

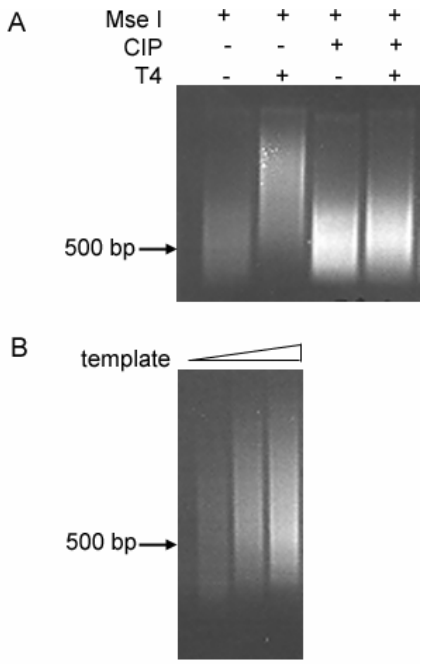

Figure 4-8. Quality controls during the preparation of the genomic DNA for MAP. Representative examples are shown. (A) Assesment of the dephosporylation reaction. Genomic DNA, that had been digested with Mse I, was dephosphorylated (CIP) and then self-ligated (T4). Positive and negative controls without dephosphorylation and/or ligation are included. The self-ligated dephosphorylated sample is indistinguishable from the non-ligated one. (B) Evaluation of the adaptor ligation efficiency. Increasing amounts of adaptor-ligated template were PCR amplified.

4.3.4. Affinity chromatography

The methylated CpG island fraction of the prepared DNA samples was affinity-purified with the MBD-column. A representative example is shown in Figure 4-9. In this example the source of the DNA was of male karyotype, thus the X-linked CpG islands —such as that of the *HPRT* gene— are present as only one unmethylated copy per cell. The only exception is the CpG island of *Xist* that is expected to be methylated in the active X chromosome. Furthermore, the typical CpG island of the *Ccne* gene is unmethylated and the imprinted region of the *Igf2r* gene is present as both methylated (imprinted) and unmethylated DNA.
The sample was injected into the column and the unbound DNA was washed with 0.1 M salt (absorbance peak between 0 and 6 ml, Figure 4-9, A). Then, the stringency of the elution was gradually increased by increasing the salt concentration from 0.1 to a final 1M. The progress of the elution was monitored by PCR with the

diagnostic primers for the *Ccne*, *Xist*, *HPRT* and *Igf2r* CpG islands. PCR amplification of the *Ccne* and *HPRT* CpG islands showed that the non-specifically bound, unmethylated DNA eluted at salt concentrations between 0.25 and 0.8 M with a peak at 0.5 M (elution volumes 12, 36 and 24 ml respectively). The elution of the specifically bound, methylated *Xist* CpG island on the other hand started at 0.4 M and continued after the final 1 M salt concentration was reached (elution volumes 18 and 48 ml respectively). The elution peak for the methylated *Xist* was at 0.8 M (39 ml). Finally the *Igf2r* differentially methylated CpG island had, as expected, two elution peaks; the unmethylated allele peaked at 0.5 M (24 ml) together with the other unmethylated CpG islands, while the methylated allele peaked at 0.8 M (39 ml) together with *Xist*. The elution profile of total DNA during the run (absorption at 260 nm) was used to evaluate if the peaks observed after PCR amplification were specific. As it can be seen at the plot (Figure 4-9, A, top) the DNA elution during the run was continuous and the observed peaks did not correspond to elution of bulk DNA from the column at the given salt concentrations.

During the first separation, the elution spectra of the unmethylated and methylated CpG islands overlapped. In order to obtain pure, methylated CpG islands, the fractions from the first elution that contained mainly methylated DNA and only the tails of the unmethylated peak (33 to 48 ml in this example) were pooled and passed through the column for a second time (Figure 4-9, B). The second elution was performed in two steps. The first step consisted of a non-specific elution of the contaminating unmethylated CpG islands at low salt concentrations (elution volumes12-19 ml and 0.6-0.8 M salt). The second step included a gradual increase from the lowest salt concentration that eluted only methylated CpG islands (0.8M) to the maximum concentration of 1 M and then a long wash at the highest salt concentration. The two steps were separated by a long wash at the highest concentration that elutes unmethylated but not the bulk of methylated CpG islands (19-33 ml, 0.8 M). As the PCR analysis of the fractions showed (Figure 4-9, B, bottom) this second affinity purification separated perfectly the methylated CpG islands. These fractions that contained pure, methylated CpG islands were pooled and used for microarray hybridisation.

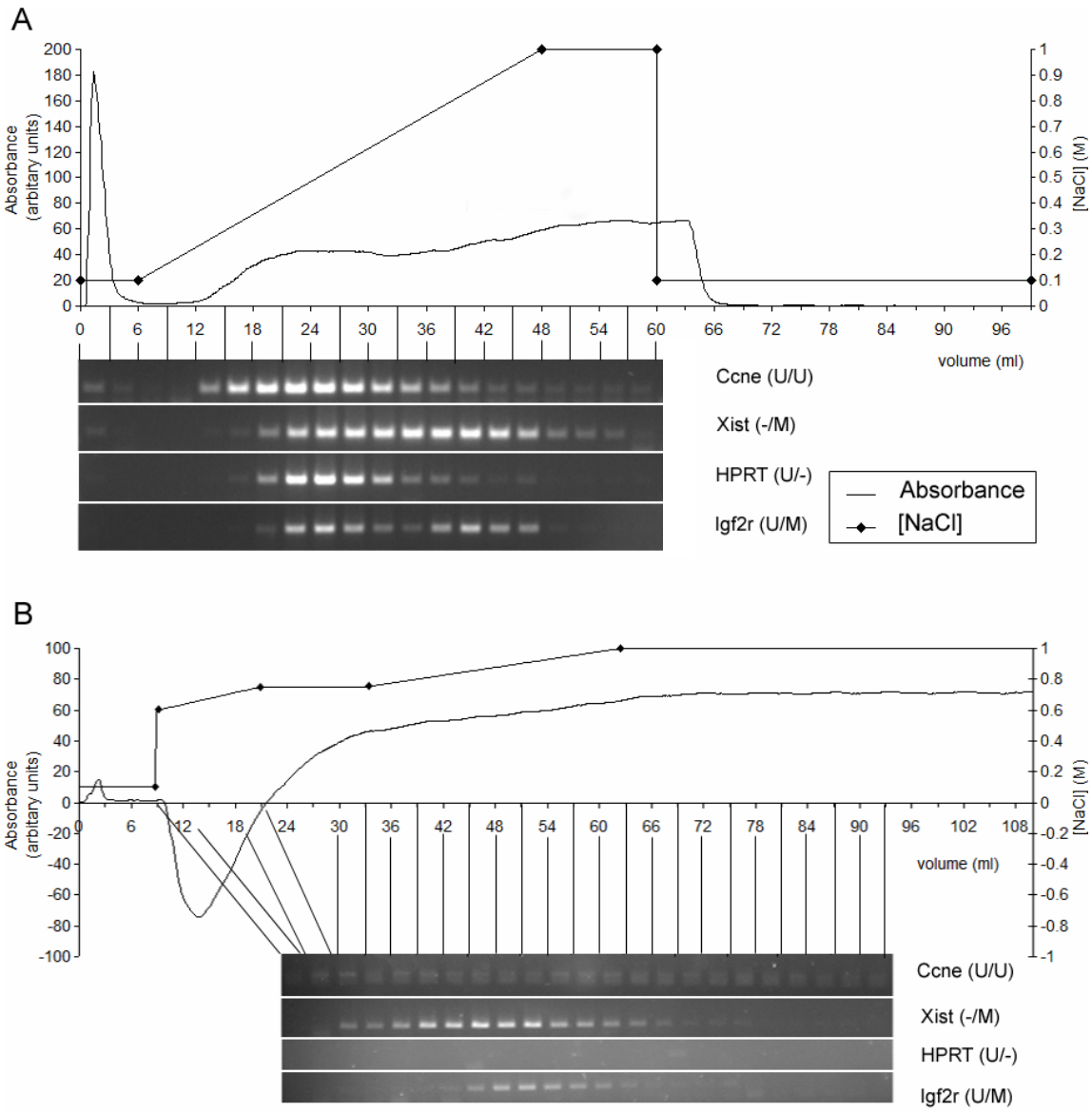

Figure 4-9. Illustration of the MAP procedure. (A) First round of affinity purification of methylated CpG islands. The elution profile of DNA, as measured by arbitrary absorbance (260 nm) units during the run is shown on top. The corresponding salt (NaCl) concentration at each point of the purification process is plotted over the elution profile chromatogram. A sample from each fraction was PCR amplified and the PCR product is shown below the corresponding fraction of the elution profile. The amplified CpG island and the methylation status of the two alleles are shown on the right of each PCR . (B) Second round of affinity purification of methylated CpG islands. As for (A).

4.4. Quality control of brain MAP and pre-processing of brain data

The brains of three mice, either female or male, were pooled together and the DNA was prepared as described previously (section 4.1 of this chapter and Materials and Methods). The DNA samples were PCR amplified with a primer specific to the adaptors and hybridised to the custom-made CpG island oligonucleotide arrays. The methyl-CpG island enriched fractions after MAP were hybridised against the input DNA of each sample for the elimination of possible bias during the preparation of the DNA or during the hybridisation. This process was repeated three times for each sex, so in total the DNA from nine female and nine male mice was hybridised on six microarrays, three for each sex.

4.4.1. The microarray data show methylation of CpG islands on the inactive X chromosome

To verify that the hybridisation worked and that the post-hybridisation, normalisation process had not introduced unacceptable biases to the results, the signal ratio of enriched versus input DNA of all the "female" arrays was plotted against that of the "male" arrays (Figure 4-10). In such comparisons, the higher the values in each axis, the higher the enrichment of the particular CpG island relative to the input. The two sources of DNA are not expected to have any major differences in their methylation patterns except for those of the inactive X chromosome. Indeed, the signal ratios generally aligned on the diagonal of the plot, showing equal enrichment in both sources. The only difference between female- and male-derived DNA was a "spike" towards higher values in the female axis. Further analysis showed that this "spike" consisted entirely of probes specific for the X chromosome (Figure 4-10, A). It would not be unreasonable to assume that this "spike" correctly showed the increased methylation levels in the CpG islands of the inactive X chromosome in females. As a control, the signal ratio of the spots that corresponded to chromosome 16 were plotted (Figure 4-10, B) and they all aligned on the diagonal with the bulk of the autosomal CpG islands, as expected.

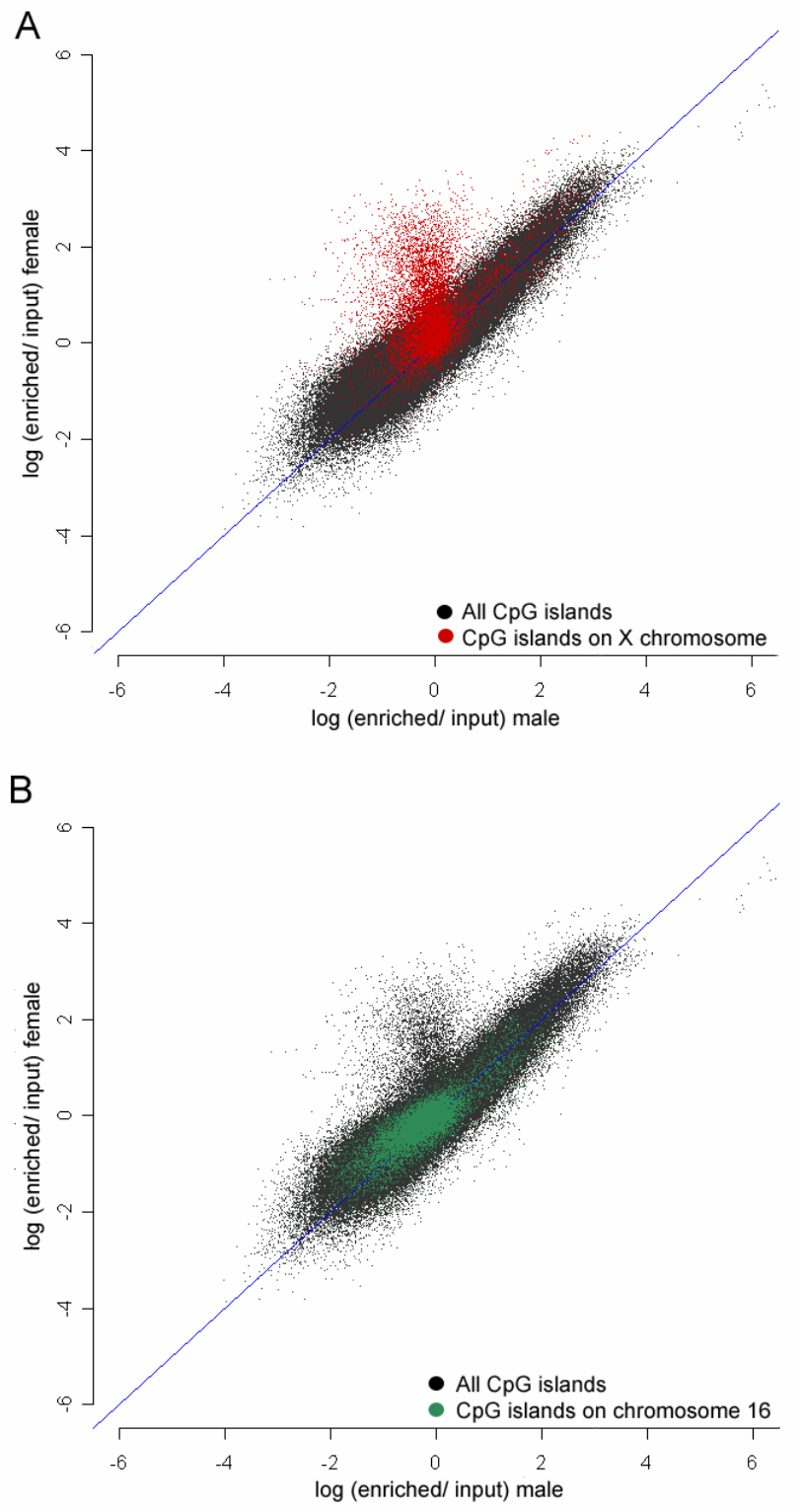

Figure 4-10. Female versus male scatter plots of the microarray data. The logarithm to base 2 of the MAP-enriched versus input signal ratio for each probe of the female and male samples is plotted on the X- and Y-axis respectively. The probes that correspond to CpG islands of the X chromosome are plotted in red (A) and those that correspond to chromosome 16 are plotted in green (B).

4.4.2. Calculation of the signal threshold for enriched Mse I fragments

There are technical issues that prevent direct estimation of the absolute enrichment levels from normalised microarray data. This means that a normalised logarithm of signal ratio (henceforth referred to as the M value) equal to one does not necessarily mean two-fold enrichment. A threshold of the normalised microarray data that signified enrichment of the particular genomic fragment is therefore needed. Second, as described previously (section 4.3.1), every Mse I fragment was represented on the array by many, tiled oligonucleotides. As a consequence of that, the enrichment of each fragment was described by multiple M values, one for each oligonucleotide. In some cases the M values for one Mse I fragment were very similar, but in others they were distributed over a larger range of values. Because of the preparation method of the DNA however, local intensity peaks within one Mse I fragment were not expected even if the local concentration of methylcytosines within the fragment varied. This meant that a summarisation method was needed for the microarray data of each CpG island, which would reflect its enrichment in the sample.

Quantitative real-time PCR (qPCR) was employed to find the threshold at which an Mse I genomic fragment could be considered as enriched. In more detail, twenty-four fragments were chosen from the array so that they represented a variety of M values (Figure 4-11). Specific primers were designed for each Mse I fragment and qPCR was performed in triplicate on equal amounts of MAP-enriched and input DNA for each primer pair. The pooled DNA from six female and three male brains was used in each PCR. By using equal amounts of DNA and not relying on the overall concentration of DNA in the input and MAP-enriched samples, problems associated with DNA recovery efficiency and amplification efficiency were minimised.

The ratio of the MAP-enriched versus input sample of each Mse I fragment tested, as determined by qPCR, showed a very good positive correlation with its mean M value from the microarray (Pearson's correlation 0.77, p=1.86 10-5) (Figure 4-12, A). In order to calculate the threshold M value, the ratios obtained by qPCR

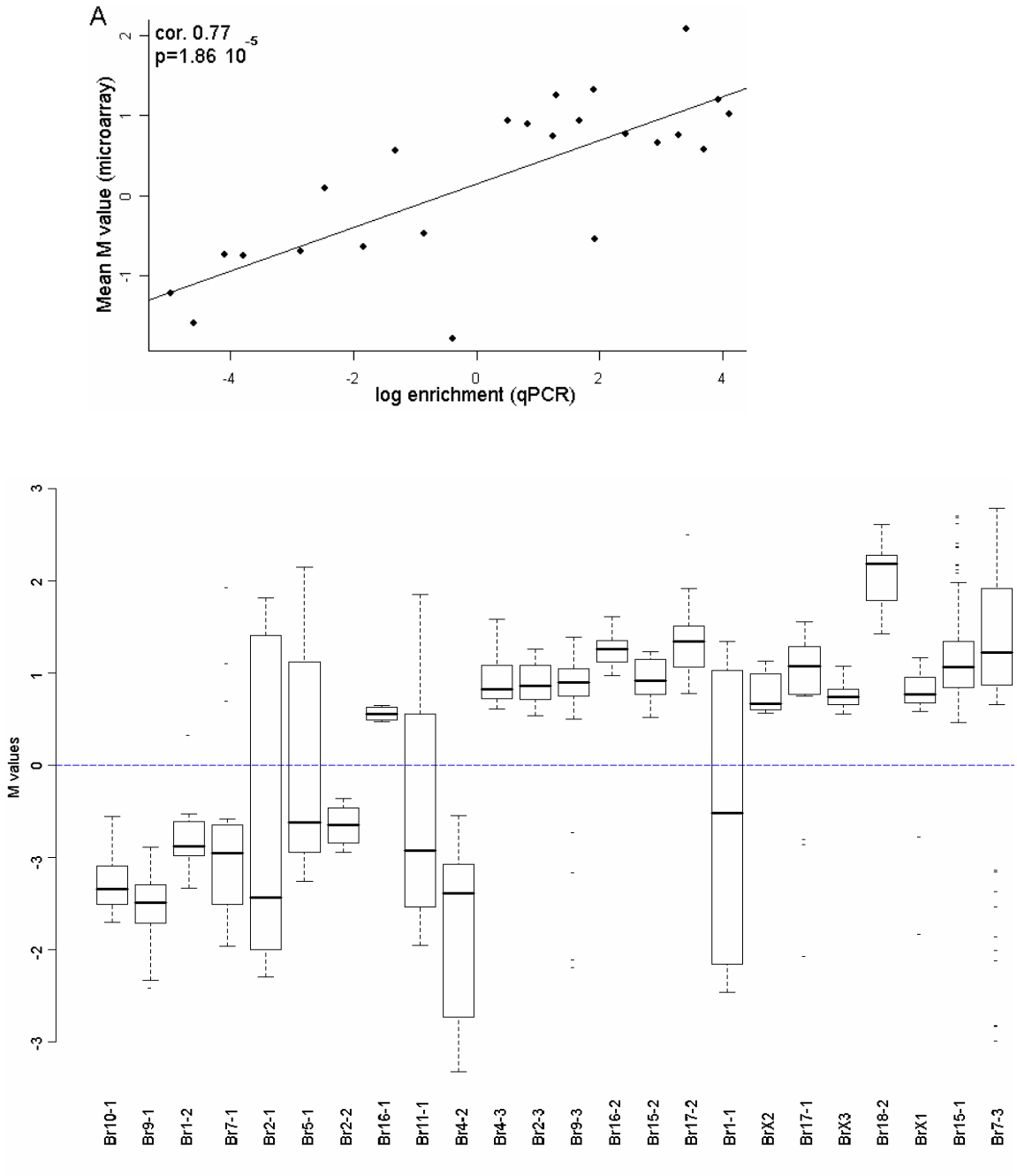

Figure 4-11. Distribution of the M values of the 24 Mse I fragments used for the calculation of the signal threshold. The M values are the normalised logarithm to base two of the signal ratio of the MAP-enriched DNA versus input. The distribution of the M values of each Mse I fragment is shown as a boxplot in which the bottom and top of the box are the first and third quartiles and the horizontal line inside the box shows the median. The whiskers in each plot extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box and the outliers of the distribution are shown with dashes above and below each plot. The identifiers of the CpG islands are shown below each boxplot.

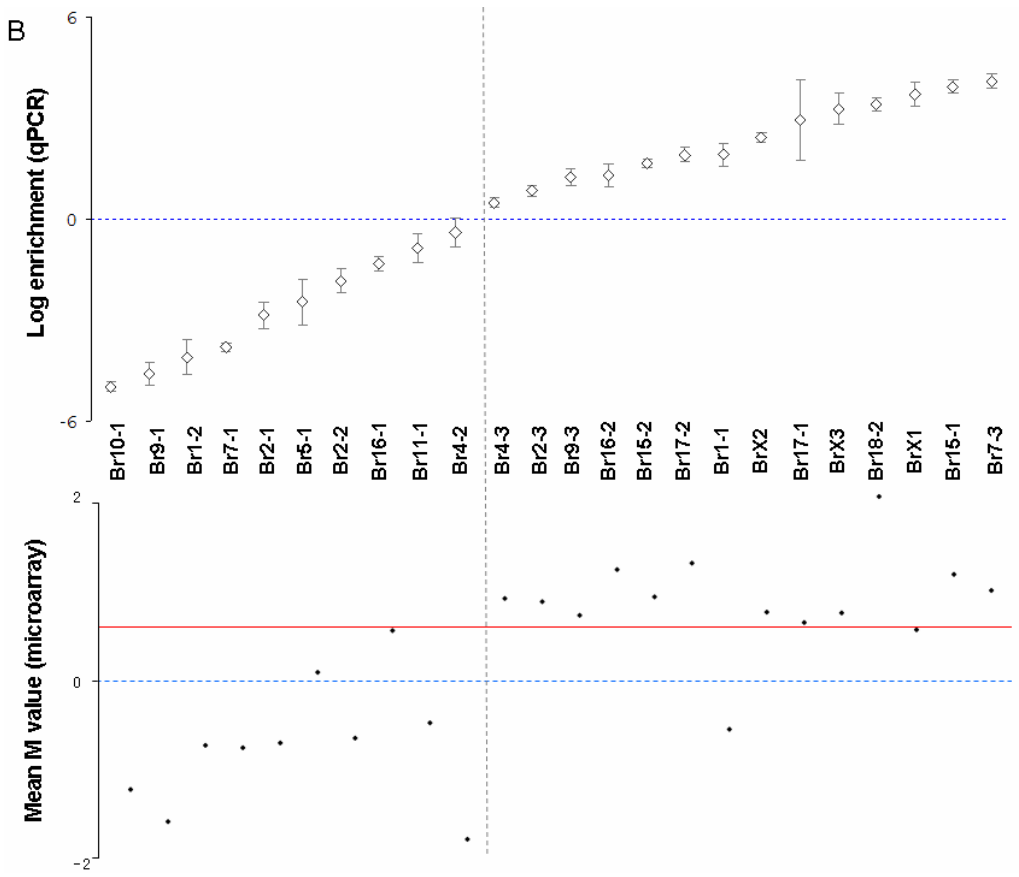

Figure 4-12. Calculation of the signal threshold for enriched CpG islands. (A) Correlation plot of the relative enrichment of each CpG island as calculated by qPCR and microarray hybridisation. The calculated Pearson's correlation of the data and the confidence value (p) is given on the top of the plot. (B) Plots used for the calculation of the threshold. The logarithms of the MAP/ Input values as calculated by qPCR were plotted in ascending order and the boundary between enriched and non-enriched CpG islands was drawn (dashed vertical line). The error bars show the technical variation. Next, the minimum M value that excluded all the points on the left of the boundary was found (horizontal red line) and was used as the threshold.

were plotted in ascending order and the fragment that showed the lowest positive enrichment value was determined (fragment Br4-3, Figure 4-12, B). This fragment was then used for the determination of the threshold M value. In more detail, the mean M values of the 24 fragments were plotted in the same ascending order as determined by qPCR. The fragments that were placed left of fragment Br4-3 were all considered as having M values that represent no enrichment and the lowest M value that excluded all of them from the sample was determined. According to this analysis

the threshold mean M value for the enriched CpG islands was 0.6 (1.5 fold) and it excluded all the CpG islands that were not enriched according to the qPCR (false positives) and only 2 (14%) of the enriched ones (false negatives). This meant that the threshold of M=0.6 was stringent in that it produced a low percentage of false negatives but did not allow for false positives. Moreover, when the distribution of the mean M values of all the Mse I fragments on the array was plotted (Figure 4-13), there were two distinct peaks. The peak at the left consisted of unenriched fragments of low M values and the peak at the right of the fragments with higher M values that are enriched after MAP. Importantly, a mean M value of 0.6 lays just before the peak of enriched fragments.

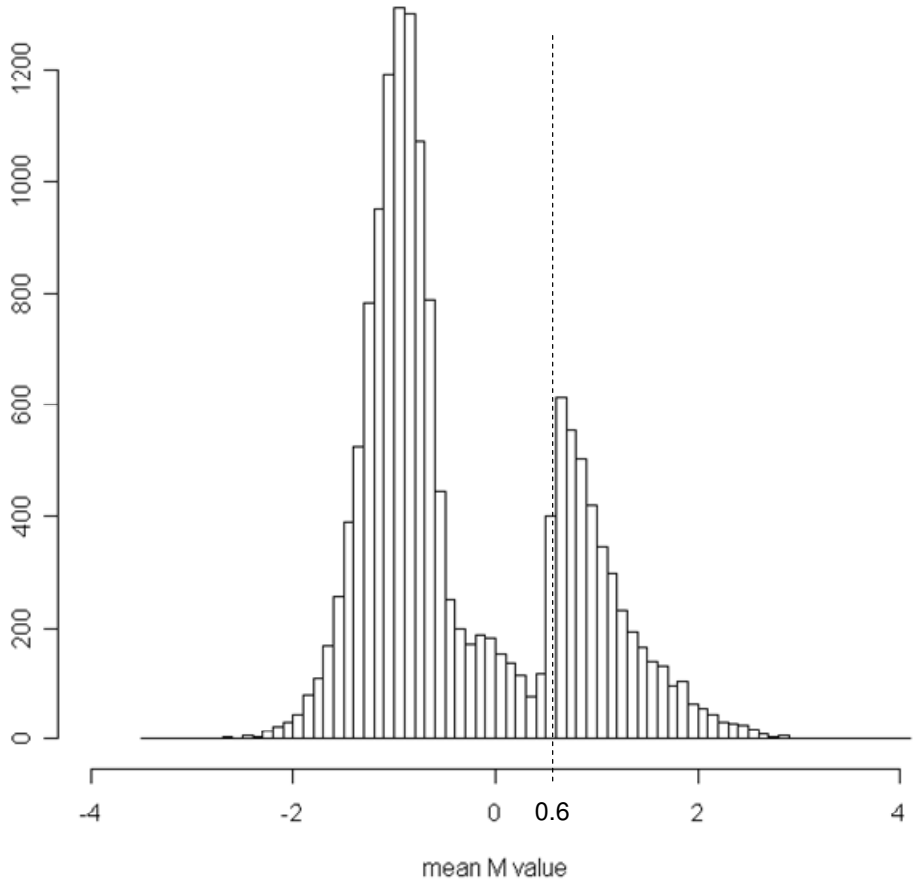

Figure 4-13. Distribution of the mean M values of all the Mse I fragments on the microarray after hybridisation with brain DNA. The x axis shows the mean M value and the y axis shows the number of Mse I fragments. The dashed vertical line shows mean M value equal to 0.6.

To further validate the use of the M=0.6 threshold, the M value distributions together with the mean M value of Mse I fragments of known methylation status were plotted (Figure 4-14). The Mse I fragments that consisted the differentially methylated regions (DMRs) of *Xist* and various imprinted CpG islands were used as positive controls (Figure 4-14, A). As expected, the range of M values for these

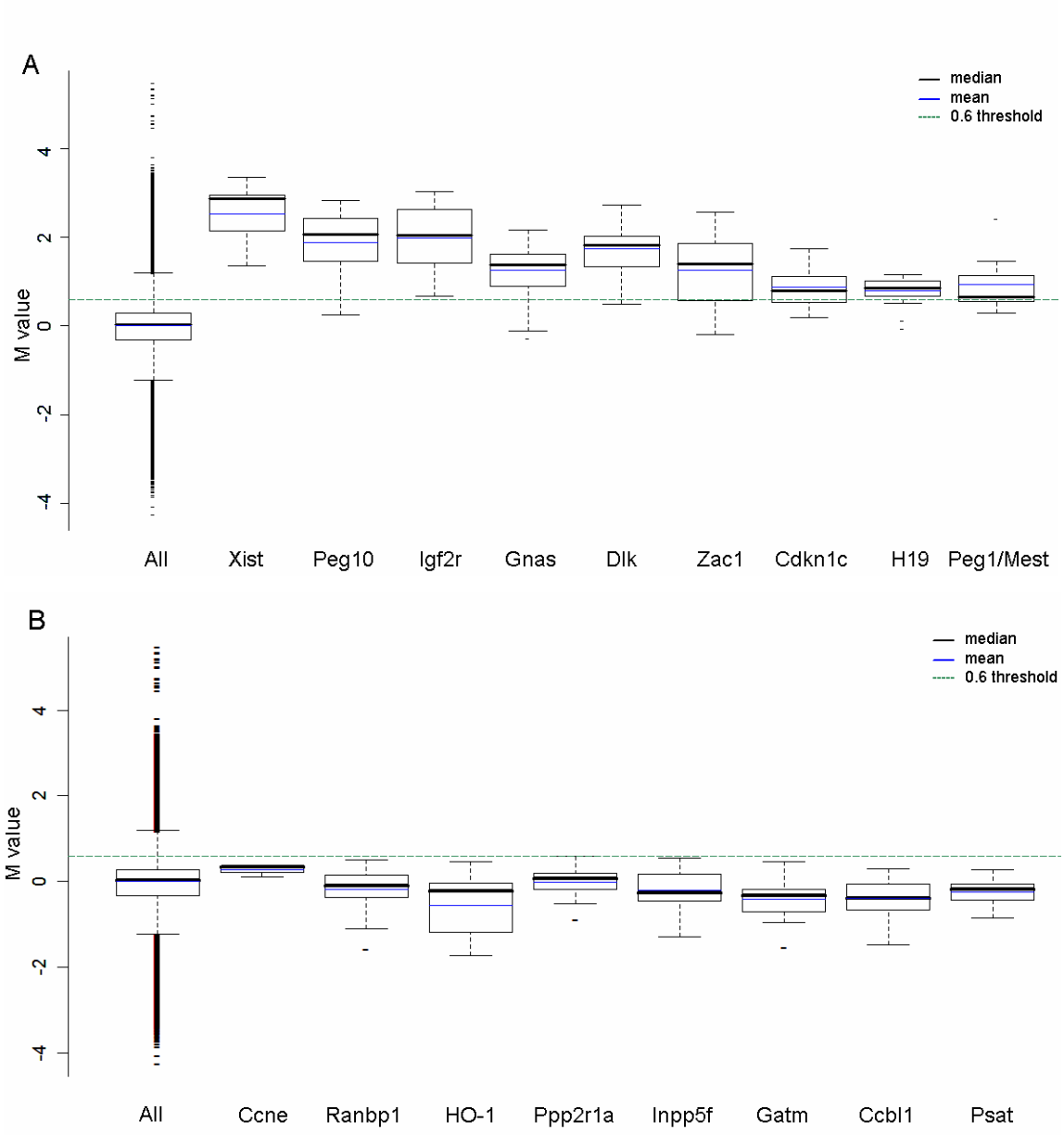

Figure 4-14. Application of the M=0.6 threshold on Mse I fragments that contain CpG islands of known methylation status. (A) M value distribution of methylated CpG islands. The mean M values of all the CpG islands are greater than 0.6. (B) M value distribution of unmethylated CpG islands. The mean M values of all the CpG islands are smaller than 0.6.

fragments was clearly shifted in comparison to the M values of all the probes on the array. Moreover, in all cases, the mean M value was above the threshold of 0.6. A similar analysis was performed for the CpG islands of housekeeping genes (Figure 4-14, B). These genes are expected to be free of methylation and therefore depleted in the sample. The distribution of the M values of the fragments that corresponded to these CpG islands was not noticeably different from that of the entire array. Importantly, a few of them appeared to have mean and median M values above zero. Application of the M=0.6 threshold on their mean M values however, successfully removed them from the group of enriched CpG Mse I fragments.

Finally, analysis of the CpG island of *Ddx4* (Figure 4-15) agreed with the previous report that it is densely methylated in brain (Song *et al.* 2005). During MAP, this CpG island was eluting with the methylated fraction of genomic DNA. Moreover, in the array data its average M value was 1.64, well above the M=0.6 threshold and comparable with those of *Xist* and *Peg10* (2.5 and 1.88 respectively).

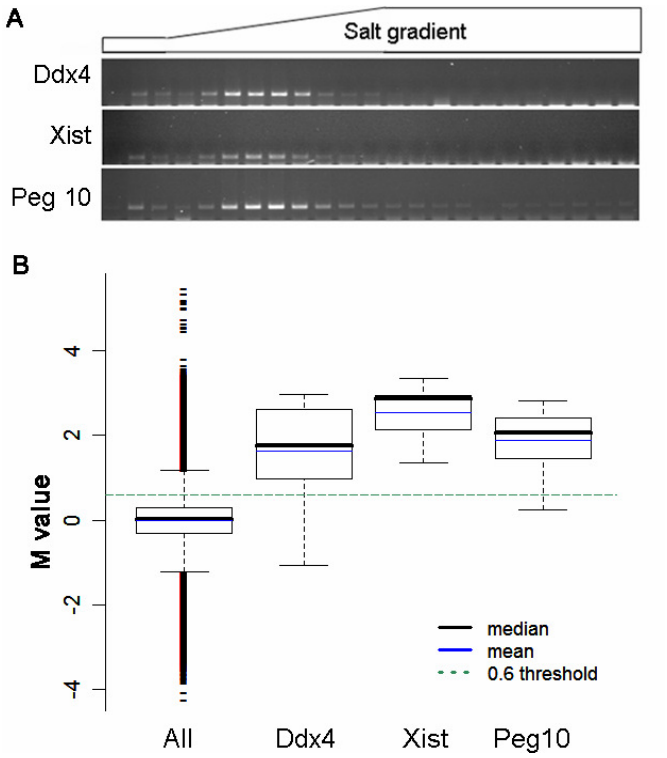

Figure 4-15. Enrichment of the known methylated CpG island of *Ddx4*. (A) In the second elution during MAP, Ddx4 eluted in the methylated fractions together with *Xist* and *Peg10*. (B) The distribution and mean M values of Ddx4 are above the 0.6 threshold and comparable to those of *Xist* and *Peg10*.

4.4.3. Methylation analysis of enriched Mse I fragments

Before proceeding with global analysis of the methylation status of CpG islands in mouse brain it was necessary to verify that the subset of Mse I fragments that passed the M=0.6 threshold were indeed significantly methylated. In other words, there is no information on how the enrichment of each fragment after MAP relative to the input correlates with its degree of methylation. Or, how methylated is a genomic fragment that has a mean M value of 0.6 and what, for instance, is the difference with a fragment that has a mean M value of 4?

To gain an understanding of how the mean M values correlated with the degree of methylation, regions of three genes were chosen to be analysed in detail by bisulfite genomic sequencing. These regions corresponded to the CpG islands of the genes *Celsr1*, *Celsr2* and *Celsr3*. These genes are all very CpG-rich and are represented on the array by many sequences/Mse I fragments, derived from both the promoter/first exon, as well as from regions of the gene body (Figure 4-16). As it can

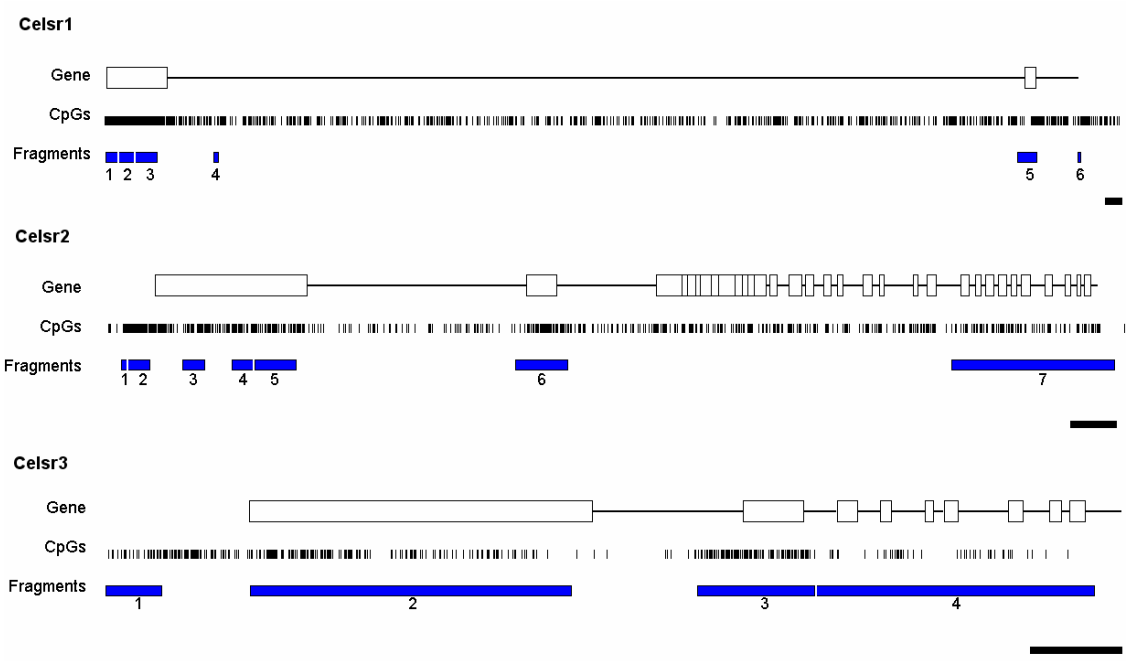

Figure 4-16. Diagram of the Celsr regions that are present on the array. In each case the gene structure is shown on top with a 5' to 3' prime orientation, starting from the first exon. Empty boxes represent exons. The vertical lines underneath the gene cartoon correspond to the positions of CpGs. The numbered blue boxes at the bottom indicate the regions that are present on the array. The left and right ends of each of these boxes signify Mse I positions. The filled black boxes at the bottom right represent a length of 1 Kb.

be seen in Figure 4-17, according to the microarray data the different regions of these genes showed different enrichments.

The mean M values of all the fragments of *Celsr1*, except for the one derived from the 5' end, were above the threshold (Figure 4-17, A). According to this, fragments 2, 3, 4, 5 and 6 should all be methylated. However, it was obvious that the intensity of the hybridisation was not equally high for all the methylated fragments. The mean M value of fragment 3 for example was twice that of fragment 2. Bisulfite genomic sequencing showed that this difference between the two mean M values had a biological meaning (Figure 4-18, A). It looks like fragment 2 is largely unmethylated and that the densely methylated last few CpGs were responsible for carrying it to the enriched fraction during MAP. Fragment 3 on the other hand, gave the impression of being more uniformly methylated. It appears that the local high concentration of methylcytosines at the 3' end of fragment 2 was responsible for raising the mean M value of the fragment above the threshold. The low overall methylation however was reflected in the relatively low mean M value. It is interesting to observe that the transition from unmethylated to methylated happened sharply, within the 139 bp (8 CpGs) that separated amplicons 3-1 and 3-2.

According to the M=0.6 threshold, all the Mse I fragments that map on the *Celsr 2* gene were enriched and therefore methylated (Figure 4-17, B). Methylation analysis of fragment 5, which had a mean M value equal to 1.88, confirmed that it was densely methylated (Figure 4-18B). Fragments 6 and 7 were marginally above the M=0.6 threshold (mean M values 0.61 and 0.66 respectively). It was interesting to investigate if in these cases the M=0.6 threshold was correct in including these fragments in the methylated CpG island pool. Genomic bisulfite sequencing of a random region of fragment 6 showed dense methylation, validating the use of the M=0.6 threshold (Figure 4-18, B).

Finally, none of the fragments that mapped on the *Celsr 3* gene passed the M=0.6 threshold (Figure 4-17, C). Fragment 2 however was very close to it, with a mean M value of 0.54. Bisulfite genomic sequencing showed that it was correct not to include this fragment in the pool of methylated CpG islands as methylation in it was very low and sporadic (Figure 4-18, C).

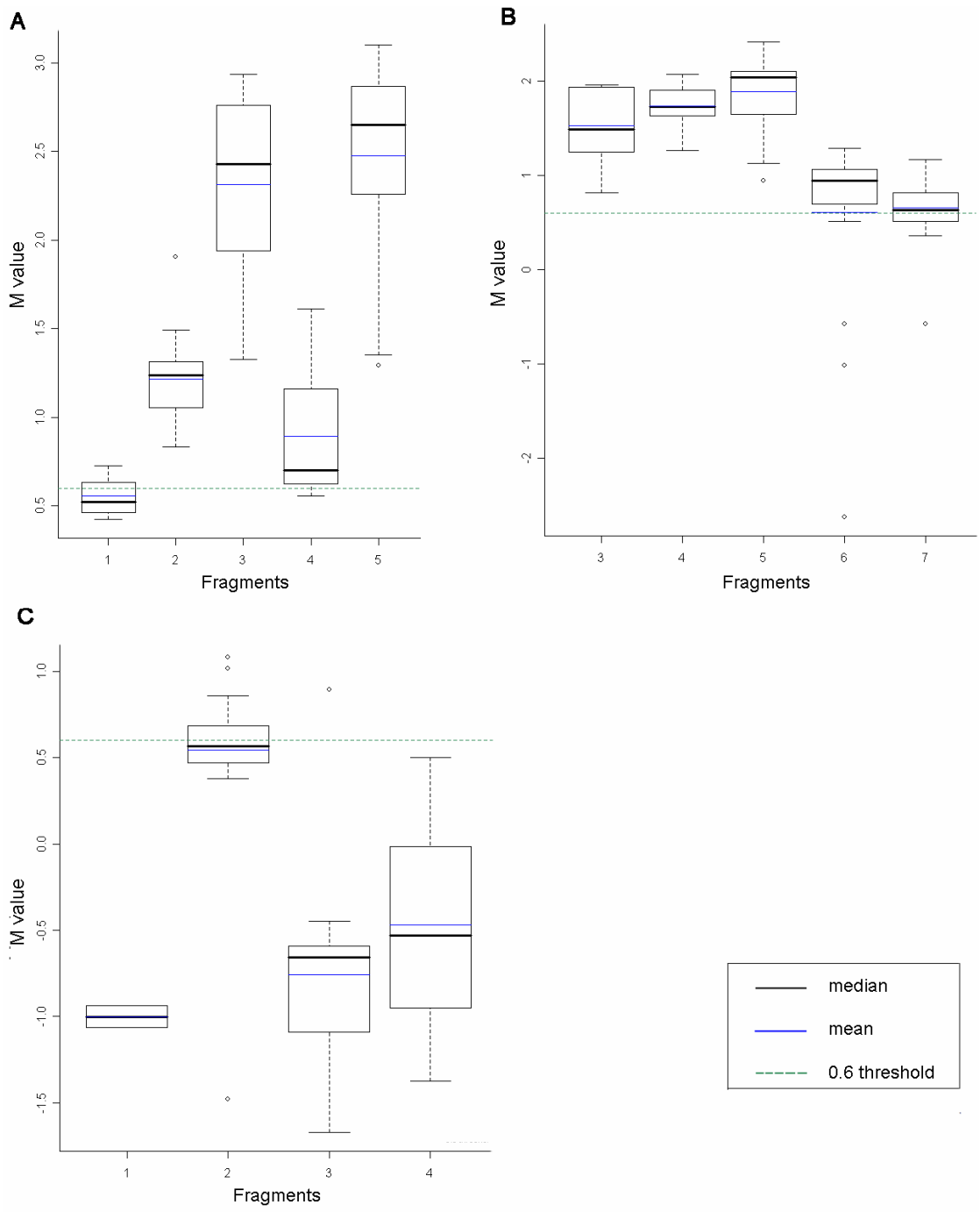

Figure 4-17. M value distributions of the different Mse I fragments/regions of *Celsr 1* (A), *Celsr 2* (B) and *Celsr 3* (C). The relative positions of the Mse I fragments/regions in the genes are shown in Figure 4-16. The fragment 6 of *Celsr 1* and the fragments 1 and 2 of *Celsr 2* are omitted because their M values were of low confidence (*i.e.* had high q values).

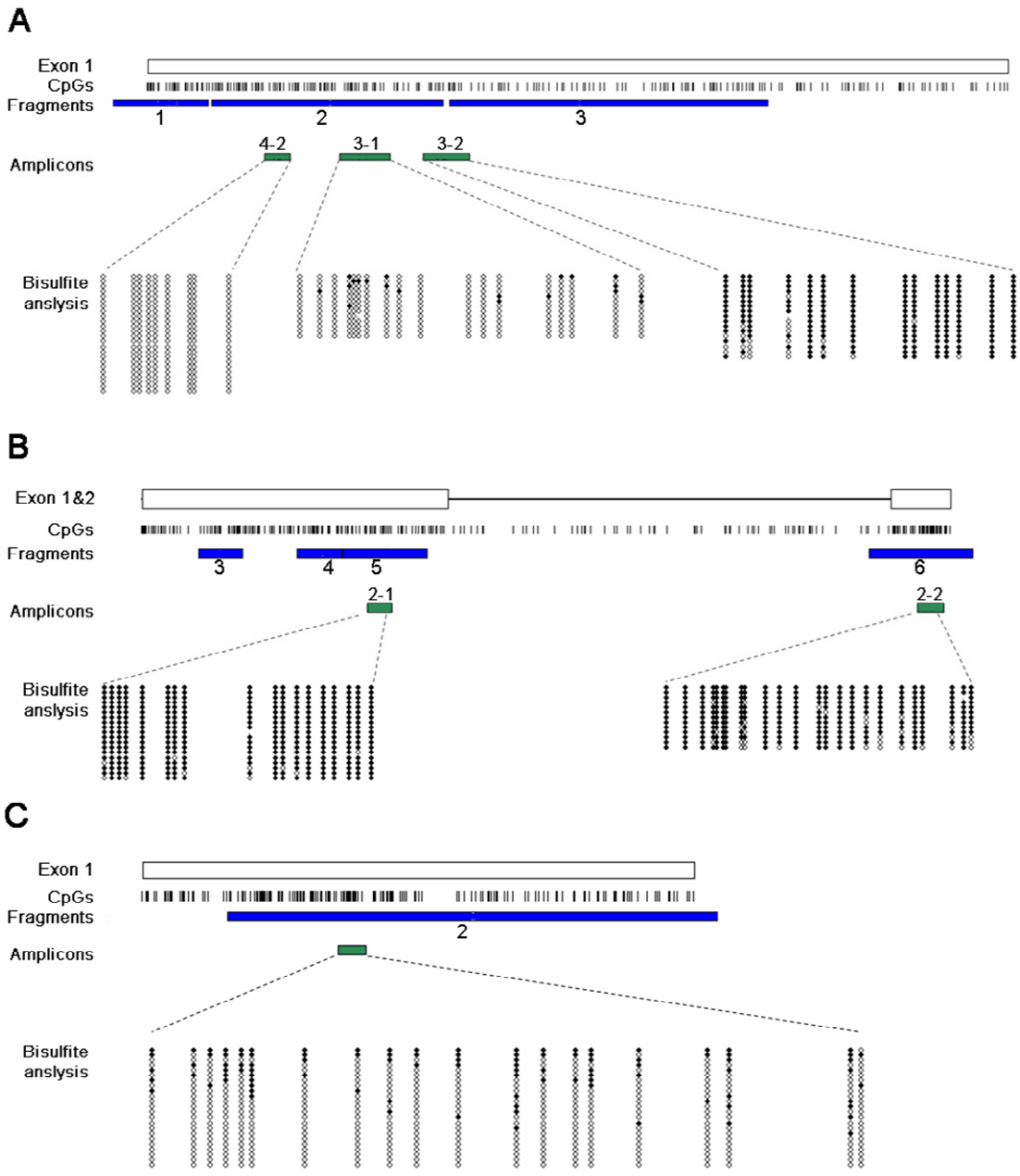

Figure 4-18. Bisulfite genomic sequencing of various CpG island regions. (A) *Celsr 1*, (B) *Celsr 2*, (C) *Celsr 3*. Close-up of Figure 4-16. The green boxes show the regions that were analysed. Filled circles are methylated CpGs and empty circles unmethylated CpGs.

In conclusion, bisulfite analysis of several fragments with various M values confirmed that the use of the M=0.6 threshold could be applied for the identification of methylated CpG islands in this experiment. Fragments with mean M values marginally above the threshold were methylated and fragments with mean M values slightly below the threshold were not. Dense but local methylation in a fragment is

enough to raise the mean M value above the threshold. On the other hand, fragments with sporadic, low methylation are not included in the pool of methylated DNA.

4.5. Global trends of methylated CpG islands in mouse brains

4.5.1. CpG density of methylated CpG islands

Similar analyses of human methylated CpG islands have shown that CpG islands of medium CpG density are preferentially methylated. In order to examine if this is the case in mouse too, all the Mse I fragments that passed the M=0.6 threshold were selected and their o/e was calculated. The means, medians and standard deviations of the o/e in both the total and the methylated fragments are shown in Table 4-3. The actual distribution of the o/e values in the enriched CpG islands in comparison to the original distribution of all the fragments is shown in Figure 4-19. In order to asses if a certain class of sequences becomes preferentially methylated in brain, three statistical tests were performed; the t-test interrogates the difference of the means, the Wilcoxon test interrogates the difference of the medians and the Kolmogorov-Smirnov (KS) test examines differences in the shape of the distributions.

By examining Figure 4-19 one can notice that the distribution of the methylated fragments is not symmetrical; there is a sharp rise at low values, a local peak at around 0.8 and a slow fading at higher values. The distribution of the entire

Table 4-3. Measures of central tendency of the o/e values in the total and methylated Mse I fragments.

	All	Methylated
Mean	0.775	0.896
Median	0.929	0.940
St.Dev.	0.446	0.327

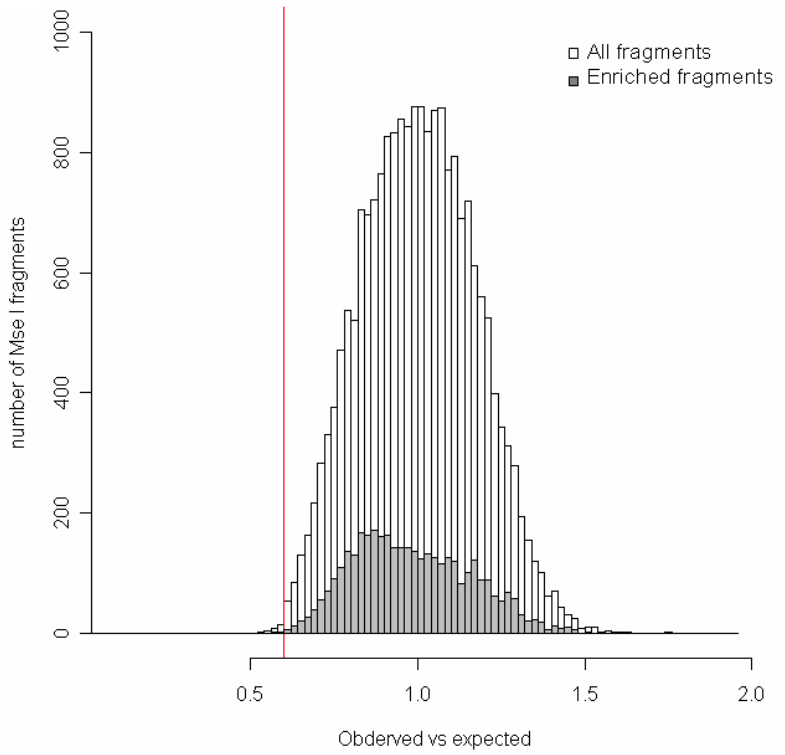

Figure 4-19. Histogram of the o/e of the methylated Mse I fragments. Grey bars are used for the histogram of the methylated fragments and white for that of all the fragments. The o/e was calculated along 200 nucleotide windows. The vertical red line shows the 0.6 threshold of the CpG island definition.

set of fragments on the other hand is symmetrical with a peak at around 1. In other words, it appears from the distribution plot that the methylated CpG islands tend to have lower o/e values than the entire set of CpG islands. Such a conclusion is supported by the KS test (p value $<2.2\ 10^{-16}$). On the other hand, the mean and themedian of the o/e values (Table 4-3) show that the methylated CpG islands are enriched for high CpG content (p values of t-test and Wilcoxon test $<2.2\ 10^{-16}$). Given these conflicting results, it is difficult to draw conclusions about a general tendency of the methylated CpG islands regarding their CpG content.

4.5.2. Distribution of methylated CpG islands in the genome

In order to analyse the global distribution of the methylated CpG islands, the CpG islands that were located on the X chromosome were excluded from the analysis. The reason for this was that the presence of the inactive X chromosome in

the sample could interfere with the interpretation of the global patterns of CpG island methylation.

18.38% of the CpG islands were methylated in brain. This corresponds to 42.77% of all the CpG island-associated genes. The observed methylation patterns according to the genomic location of the CpG island are shown in Table 4-4. It was next examined whether there was a tendency for more than one region to become methylated together in the same gene, or whether methylation was happening independently. Of the 2385 genes that containing CpG islands in both the 5' and the gene body, 11.9% showed methylation in both regions. Of the 215 genes that contained CpG islands in both 3' and the gene body, 10.23% showed methylation in both regions. Finally, of the 438 genes that had CpG islands in both the 5' and 3' regions, 7.53% showed methylation. There were no genes that had CpG islands in all

Table 4-4. Distribution of the autosomal CpG islands that are methylated in brain in the mouse genome

	Number of methylated CpG islands	Methylated CpG islands[1] (%)	Methylated CpG islands[2] (%)	Number of genes with methylated CpG islands	Genes with methylated CpG islands[3] (%)
Intergenic	692	13.91	18.6	NA[4]	NA[4]
5'	1598	15.63	43	4340	34.72
Intragenic	1340	27.77	36.03	2263	46.76
3'	207	21.3	5.56	174	23.32
Total	*3719*[5]	*18.38*	*100*[5]	6445[5]	*42.77*[5]

[1] The percentage of methylated autosomal CpG islands relative to the total autosomal CpG islands per location.

[2] The percentage of methylated autosomal CpG islands per genomic region relative to all the methylated autosomal CpG islands.

[3] The number of genes that are associated with the methylated autosomal CpG islands relative to all the genes that are associated with CpG islands in the indicated location.

[4] Not applicable.

[5] This number is smaller than the sum of the rows above because of the cases of CpG islands mapping in more than one gene locations.

three regions. It seems like the different regions of the gene are in principle acquiring methylation independently from each other.

It was previously shown that 1065 CpG islands (1047 if only the autosomes are considered) overlap with the 5' regions of two genes with opposite orientations in the genome (section 4.2). This introduces the interesting possibility that the two genes share a common regulatory mechanism and the presence of a CpG island in these cases might have a special significance. In order to investigate this scenario, the methylation status of these CpG islands was examined. 145 of these CpG islands (13.8%) were methylated. When this number is compared to the total fraction of 5'CpG islands that become methylated (15.63%, Table 4-4), it seems like there is no particular tendency of this category of 5' CpG islands to be methylated more or less often than the rest.

4.5.3. Gene ontology of the methylated genes

In order to identify general methylation trends of the genes that are associated with CpG island genes in brain, gene ontology analysis was performed. The genes that were associated with methylated CpG islands were compared against all the CpG island-associated genes using the PANTHER interface (Thomas *et al.* 2003; Mi *et al.* 2005, http://www.pantherdb.org) and Bonferroni correction for multiple testing. The results of this analysis are shown in Table 4-5.

It becomes immediately evident that the genes that play a direct role in development and were seen to be preferentially associated with CpG islands, are also showing preferential methylation. Moreover, genes that participate in the cadherin and Wnt signalling pathways that also have a role in development are also preferentially methylated. The *Celsr* genes, whose methylation pattern was analysed in detail in section 4.4.3, belong to the cadherin superfamily. The significance of methylation in the CpG islands of genes involved in cell structure and communication and the cytoskeleton is not immediately obvious and could involve tissue-specific methylation events.

The distribution of methylated CpG islands in the different gene regions is important for the interpretation of the methylation data. It is generally considered that methylation at the 5'end of a gene is related to transcriptional inactivity, something that is not as widely accepted for the gene body and 3' regions. Moreover, it has been

Table 4-5. Gene ontology classification of the genes that are associated with methylated CpG islands.

GO term	No. of genes with CpGi	No. of genes with methylated CpGi	Expected[4]	P value
Developmental processes (B.P.) [1]	1040	263	207.4	$1.17\ 10^{-3}$
Cadherin (M.F.) [2]	43	23	8.58	$4.96\ 10^{-3}$
Cadherin signaling pathway (P.) [3]	79	39	15.75	$6.86\ 10^{-5}$
Wnt signaling pathway (P.) [3]	175	64	34.9	$6.90\ 10^{-4}$
Cell structure and motility (B.P.) [1]	459	142	91.54	$8.48\ 10^{-6}$
Cytoskeletal protein (M.F.) [2]	290	83	57.83	$2.55\ 10^{-2}$
Cell communication (B.P.) [1]	525	141	104.7	$3.92\ 10^{-2}$

[1] B.P., biological process, [2] M.F., molecular function, [3] P., pathway

[4] Number of genes in this category expected to be associated with methylated CpG islands by chance.

suggested that gene body methylation might be specific to housekeeping genes (Suzuki *et al.* 2007). In order to examine the existence of any such trend, the same analysis was performed separately for the 5', 3' and gene body-associated CpG islands.

It is expected that methylated CpG islands at the 5' will associate with genes that are not important for brain function. Indeed, as Table 4-6 shows, this is generally the case. However, it appears that the categories of genes that are preferentially associated with 5' CpG island methylation are not as diverse as the highly specialised brain function would indicate. In particular chemosensory genes and genes that are specific for the nasal cavity (olfaction and pheromone response) are the main categories of the methyl-CpG island-associated genes. G protein signalling is also very important for these functions. Receptors were a class of genes that was significantly correlated with CpG islands (Table 4-2). It is possible that the presence

Table 4-6. Gene ontology classification of the genes that are associated with methylated CpG islands at their 5'.

GO term	No. of genes with CpGi	No. of genes with methylated CpGi	Expected[3]	P value
G-protein coupled receptor (M.F.) [1]	319	237	102.55	$1.23\ 10^{-29}$
G-protein mediated signalling (B.P.) [2]	409	266	131.48	$1.06\ 10^{-24}$
Receptor (M.F.) [1]	596	325	191.59	$3.41\ 10^{-19}$
Sensory perception (B.P.) [2]	302	242	97.08	$6.49\ 10^{-36}$
Chemosensory perception (B.P.) [2]	158	155	50.79	$7.31\ 10^{-31}$
Olfaction (B.P.) [2]	155	153	49.83	$1.36\ 10^{-30}$
Pheromone response (B.P.) [2]	47	46	15.11	$1.50\ 10^{-8}$
Cell surface receptor mediated signal transduction (B.P.) [2]	683	356	219.56	$1.98\ 10^{-17}$
Signal transduction (B.P.) [2]	1290	532	414.69	$9.81\ 10^{-9}$
Serine protease (M.F.) [1]	68	43	21.86	$6.05\ 10^{-3}$

[1] M.F., molecular function, [2] B.P., biological process

[3] Number of genes in this category expected to be associated with methylated CpG islands by chance.

of CpG islands in these genes in particular has some special physiological meaning.

When the gene ontologies of genes that contain methylated CpG islands in their gene body are compared (Table 4-7) with those of the 5' CpG islands, it was surprising to notice that roughly the same functions are associated with methylation in both regions. Namely, receptor activity, signal transduction, sensory and chemosensory perception and olfaction are all enriched in this class of CpG islands.

Finally, despite the high occurrence of methylation in the 3' CpG islands (Table 4-4), methylation of this region showed no preference for any gene category.

Table 4-7. Gene ontology classification of the genes that are associated with methylated CpG islands at their gene body.

GO term	No. of genes with CpGi	No. of genes with methylated CpGi	Expected[3]	P value
Sensory perception (B.P.) [1]	148	145	49.66	$5.86\ 10^{-28}$
Chemosensory perception (B.P.) [1]	62	84	20.8	$7.51\ 10^{-24}$
Olfaction (B.P.) [1]	62	83	20.8	$4.33\ 10^{-23}$
G-protein mediated signalling (B.P.) [1]	249	165	83.54	$4.87\ 10^{-214}$
G-protein coupled receptor (M.F.) [2]	176	147	59.05	$5.76\ 10^{-21}$
Cell surface receptor mediated signal transduction (B.P.) [1]	437	224	146.62	$2.86\ 10^{-8}$
Signal transduction (B.P.) [1]	869	380	291.56	$4.86\ 10^{-7}$
Pheromone response (B.P.) [1]	18	24	6.04	$3.66\ 10^{-6}$
Receptor (M.F.) [2]	381	224	127.83	$9.68\ 10^{-15}$
KRAB box transcription factor (M.F.) [2]	91	66	30.53	$2.39\ 10^{-6}$
Zinc finger transcription factor (M.F.) [2]	189	101	63.41	$8.17\ 10^{-4}$

[1] B.P. biological process, [2] M.F., molecular function

[3] Number of genes in this category expected to be associated with methylated CpG islands by chance.

4.6. Detection of CpG island methylation establishment during development

In order to investigate further the relationship of CpG island methylation with development, the MAP methodology was used for the comparison of the CpG island methylation status between ES cells and embryoid bodies, their *in vitro* differentiated derivatives. The embryoid bodies were obtained from wild-type ES cells (E14) by LIF removal for three days, followed by addition of retinoic acid for ten days (RA10). DNA samples from the ES cells and the embryoid bodies were prepared, affinity purified and hybridised on the CpG island microarrays against input DNA as described (section 4.3). The results presented here are the sum of three independent differentiation experiments.

4.6.1. Validation of the microarray data

The microarrays were normalised as described in Materials and methods. Comparison of the M values (normalised logarithm of the signal ratio between the MAP-enriched sample and input DNA) between experiments showed the reproducibility was poor (Figure 4-20). This prohibited genome-wide analysis of these microarray data in a way similar to that applied in the brain data (section 4.4). The low reproducibility most probably indicated the presence of noise at the arrays due to low overall methylation levels in these cells in combination with PCR amplification as it will be discussed in more detail later.

In order to investigate whether the results from the probes that showed the least variation would allow extraction of any data from these arrays, the M values of Mse I fragments of known methylation status were plotted (Figure 4-21). In more detail, the M values of imprinted CpG islands as well as the differentially methylated region of *Xist* were plotted for both the ES and the RA10 samples. The ES cell line used in this experiment was of male karyotype and, like imprinted genes, *Xist* was expected to be methylated and show high M value. As it can be seen in Figure 4-21, in all cases, the distribution of M values of these genomic regions is shifted in comparison to the M values for the bulk of CpG islands present on the array. In

conclusion, these control CpG islands are correctly and reliably identified as being methylated by the microarray results.

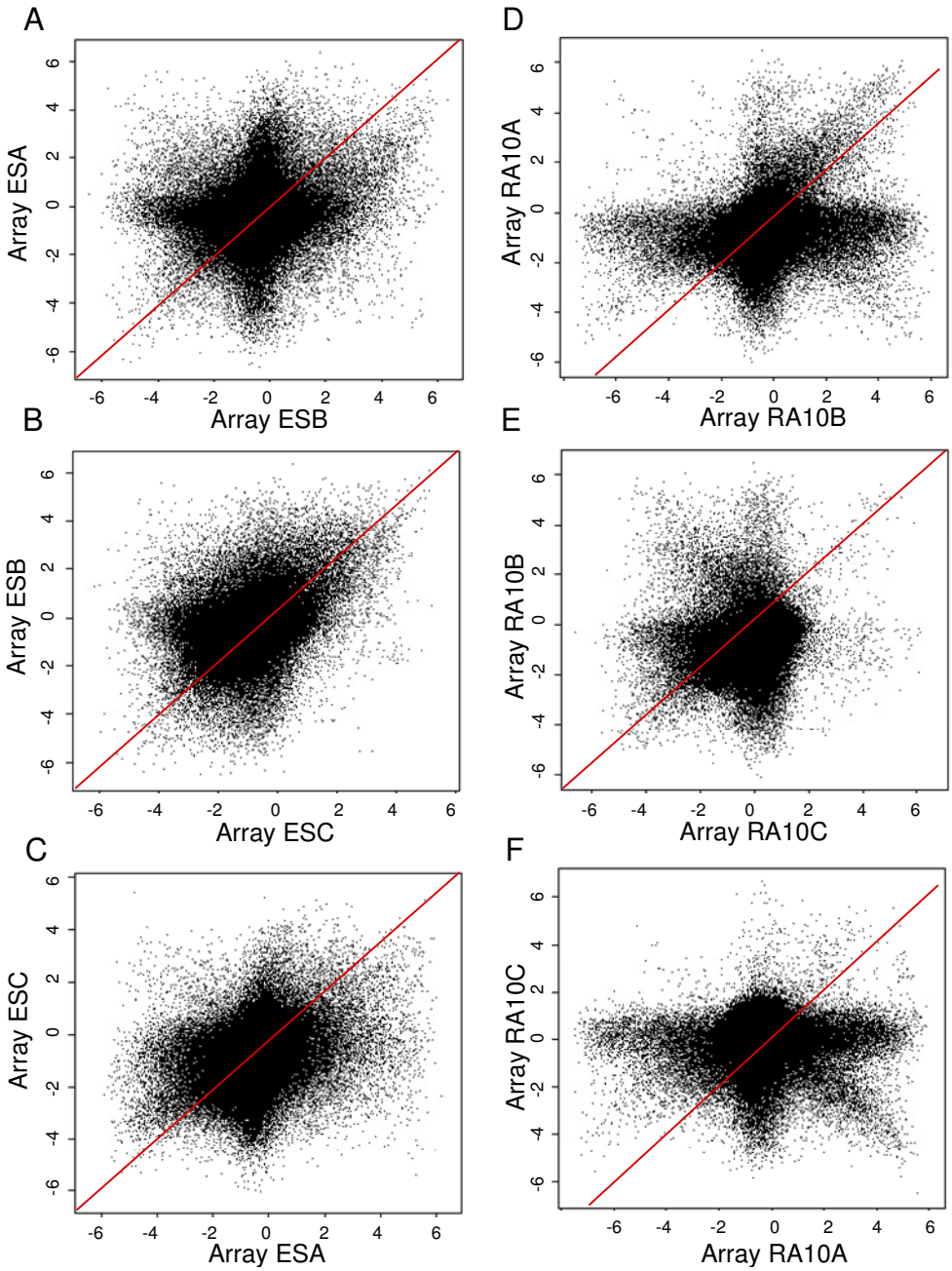

Figure 4-20. Scatterplots of the M values between replicates. The three ES arrays (Array ESA to ESC) and the three RA10 arrays (Array RA10A to RA10C) are compared. The x and y axes show the M values in the respective array. Probes that exhibit similar behaviour in both experiments lay near the diagonal. (A to C) Comparison of the microarrays that were hybridised with ES cell samples. (D to F) Comparison of the microarrays that were hybridised with RA10 samples.

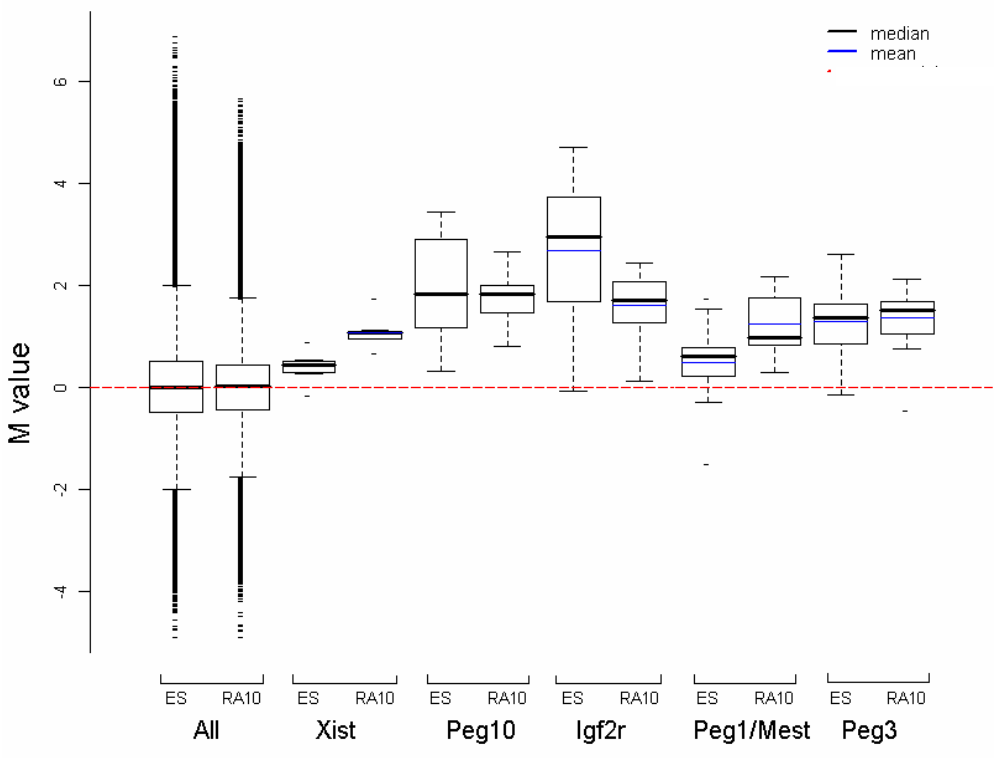

Figure 4-21. Distribution of the M values that were acquired for genomic regions that are expected to be methylated in ES cells and embryoid bodies (RA10).

Having confirmed that –at least for some of the Mse I fragments– the microarray data appeared to be trustworthy, the most consistently enriched Mse I fragments were identified. In more detail, each Mse I fragment was represented on the microarray by multiple oligos. Because the DNA preparation method during MAP involves fragmentation by Mse I digestion, the oligos derived from the same Mse I fragment should show comparable hybridisation signals. This was used for the identification of the Mse I fragments that were reliably enriched in the RA10 samples. The fragments that had an M value below or equal to 0 at the ES sample and above or equal to 2 at the RA10 sample in at least 90% of their associated oligonucleotides on the array, were selected as being potentially highly enriched. This process identified 24 Mse I fragments as being highly enriched in the embryoid bodies but not in ES cells (Table 4-8).

Table 4-8. CpG islands that are enriched in the RA10 MAP-purified samples, as indicated by microarray hybridisation.

ID	Genomic coordinates[1]	M[2] (ES)	M[2] (RA10)	Gene[3]	Location[3]	no.of exons[4]
Dev1-1	CHR1: 161132983-161133709	-0.75	1.66	C9orf40 -like	Exon 1	4
Dev1-2	CHR1: 6383009-6385694	-0.21	0.94	-	Intergenic	-
Dev2-1	CHR2: 65949836-65950273	-0.34	2.55	A33010 2K23Rik	Intron/ Exon 4	6
Dev4-1	CHR4: 122632911-122633979	-0.67	1.35	Hpcal4	Exon 3	5
Dev6-1	CHR6: 17780608-17781540	0.18	2.35	St7	Intron 4	15
Dev7-1	CHR7: 105642755-105643428	-0.42	3.41	Ascl3	Intron/ Exon 2	2
Dev7-2	CHR7: 13920979-13921968	0.46	3.07	Slc8a2	Intron/ Exon9	10
Dev8-1	CHR8: 118186994-118187994	-0.02	1.69	Hsdl1	Exon 4	5
Dev10-1	CHR10: 120756802-120757699	-0.21	1.58	493050 5D03Rik	Exon 12	12
Dev10-2	CHR10: 14224996-14225901	-0.38	0.97	Gpr126	Intron 2	25
Dev10-3	CHR10: 20124755-20125441	-0.47	3.48	Pde7b	Exon 12	12
Dev10-4	CHR10: 52362305-52362988	-0.11	2.48	Zfa	3prime	1
Dev10-5	CHR10: 59389229-59390186	-0.91	1.87	L23-like	Exon 1	1
Dev11-1	CHR11: 103521012-103522018	-1.93	1.73	Plekhm 1	Exon 1	12
Dev11-2	CHR11: 106127743-106128527	0.45	2.38	Kcnh6	5prime	7
Dev11-3	CHR11: 11720725-11721744	0.01	2.57	Ikzf1	3prime	7
Dev12-1	CHR12: 52087446-52087926	-1.27	1.64	Npas3	Intron/ Exon 12	12
Dev14-2	CHR14: 40176456-40177819	-0.33	3.52	Ptgdr	Intron/ Exon 1	2
Dev14-3	CHR14: 71932041-71933196	-0.62	1.75	Rpl23a	Exon 1	1
Dev16-1	CHR16: 3579742-3581406	-1.15	0.91	Cluap1	Exon 1	12
Dev17-1	CHR17:85056684-85057302	-1.19	2.06	Epas1	Intron/ Exon 1	16
Dev19-1	CHR19: 37393584-37394095	-0.43	2.18	EG5467 26	Exon 6	7
DevX-1	CHRX: 71110227-71110949	-0.04	2.88	C9orf40 like	Exon 1	2
DevX-2	CHRX: 89508541-89510756	0.12	0.93	Zfx	Intron 1	9

[1]According to NCBI build 34, [2]Mean M values in the indicated experiment, [3] Associated gene and location of the fragment relative to it, [4]Total number of exons in the associated gene

The enrichment of a random subgroup of these fragments was confirmed with qPCR (Figure 4-22). The qPCR was performed on equal volumes of MAP-purified DNA from ES cells and RA10. Enrichment in the RA10 samples was confirmed for eight of the twelve fragments tested with qPCR. These results, together with the previous verification that the MAP-enriched fragments are methylated (sections 4.3 and 4.4), are a good indication that the majority of the fragments in Table 4-8 are *de novo* methylated during differentiation of ES cells. However, further verification with bisulfite sequencing will be needed for the definitive confirmation of the methylation status of these genomic regions.

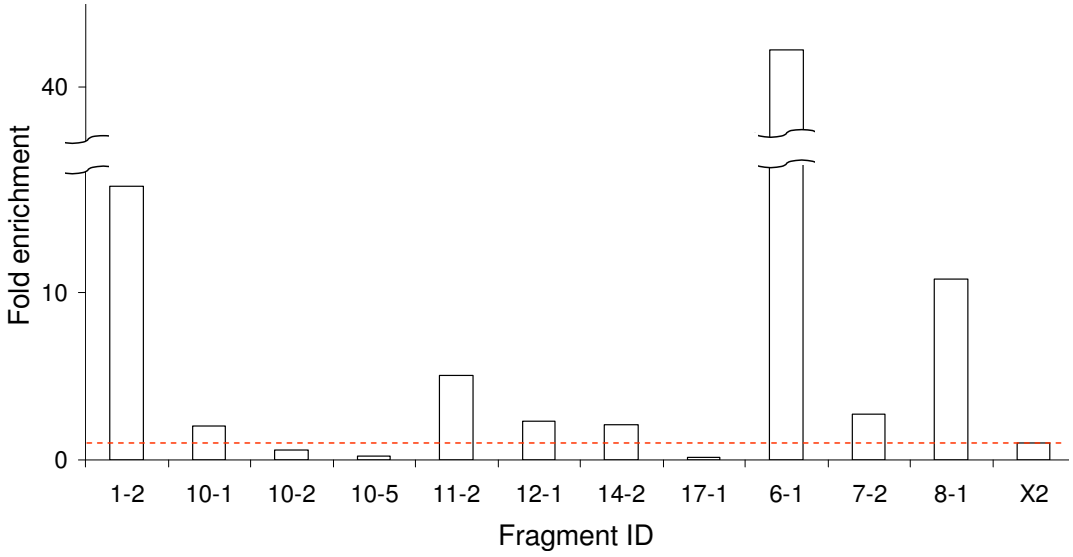

Figure 4-22. qPCR verification of the microarray data. The fragment IDs correspond to those of Table 4-8. The y axis represents the calculated ratio of the fragment in the affinity-purified DNA from embryoid bodies (RA10) relative to the affinity-purified DNA from ES cells. The red dashed line shows a ratio of 1 (no enrichment).

4.6.2. Methylation pattern of the de novo methylated fragments in brain

Fragments that appeared to be enriched in the embryoid bodies were further investigated in terms of their methylation pattern in brain. Nine out of fifteen fragments that appear to become *de novo* methylated in embryoid bodies were also methylated in brain (mean M value equal or above 0.6) (Table 4-9).

Table 4-9. Annotation of genes associated with Mse I fragments/CpG islands that become *de novo* methylated in embryoid bodies (Table 4-8).

ID[1]	M[2] (brain)	[3]	Gene	Expression pattern[4]	[5]	Biological Process[6]
Dev10-4	2.40	3p	Zfa	General	N	Unclassified
Dev7-1	2.58	Intragenic	Ascl3	Olfactory epithelium	Y	Transcription regulation
Dev10-1	**0.99**		**4930505D03Rik**	**Distinct/not brain**	**Y**	**Unclassified**
Dev10-3	1.31		Pde7b	Olfactory epithelium	Y	Signal transduction
Dev12-1	**0.49**		**Npas3**	**General**	**Y**	**Transcription regulation**
Dev2-1	1.03		A330102K23Rik	General	N	Apoptosis
Dev4-1	1.81		Hpcal4	Brain	N	Vision
Dev6-1	**2.04**		**St7**	**Zygote/egg/ umbilical cord**	**Y**	**Tumour suppressor /skeletal development**
Dev7-2	**0.02**		**Slc8a2**	**Brain**	**Y**	**Signal transduction**
Dev8-1	**1.91**		**Hsdl1**	**Distinct/ brain**	**N**	**Steroid metabolism**
Dev11-1	-0.96	5' of gene	Plekhm1	General	Y	Unclassified
Dev11-2	**2.14**		**Kcnh6**	**Distinct/not brain**	**Y**	**Signal transduction**
Dev14-2	**-0.89**		**Ptgdr**	**Dorsal root ganglia**	**Y**	**Signal transduction**
Dev14-3	-0.93		Rpl23a	Distinct/brain	Y	Protein biosynthesis
Dev16-1	-1.00		Cluap1	Distinct/brain	Y	Unclassified

Bold letters indicate that the microarray result has been confirmed by qPCR (Figure 4-22).

[1] The IDs correspond to the Mse I fragments in Table 4-8.

[2] Mean M value of the Mse I fragment in the brain microarray experiment.

[3] Position of the Mse I fragment relative to the associated gene.

[4] The expression pattern information has been acquired from the SymAtlas database (http://symatlas.gnf.org). Tissue-specific expression is indicated by the tissue in which the gene is expressed. Broader expression is characterised as “General” if there is no tissue preference and “Distinct” if there is a distinct expression pattern in different tissues. If expression is “Distinct” the expression status of the gene in the brain is especially mentioned.

[5] Y/N, the expression pattern of the associated gene agrees or disagrees, respectively, with the methylation status of the CpG island in the brain.

[6] The biological process terms were acquired from the PANTHER database (http://www.pantherdb.org).

In order to understand better these methylation patterns, the function of the genes that are associated with these fragments was explored. The gene ontology characterisation of the genes was acquired from the PANTHER database (Thomas *et al.* 2003; Mi *et al.* 2005, http://www.pantherdb.org). A few of these genes are involved in signal transduction and the list also includes transcription factors, genes involved in metabolism and apoptosis as well as genes associated with vision and skeletal development (Table 4-9). However, there did not appear to be a general trend for the genes that are associated with the *de novo* methylated fragments regarding their function.

Next, the genes were analysed in terms of their expression pattern in different tissues. The expression data were acquired from the SymAtlas expression microarray database (Su *et al.* 2002 , http://symatlas.gnf.org). A gene was considered to be expressed in a tissue if its signal was reported to be equal or more than thirty times the array median. As it can be seen in Table 4-9, the genes showed a variety of expression patterns, from tissue-specific to general expression in all the examined tissues. Importantly, the expression of a gene in brain seemed to correlate well with absence of methylation in its associated CpG island. This was true for all the genes that were associated with CpG islands at their 5' end. The methylation status of CpG islands at the 3' UTR as well as the gene body was not predicting as well the genes' expression in brain, although it did show some consistency with expression (methylation of six out of nine intergenic CpG islands was in accordance to the gene's transcription status).

4.7. Discussion

CpG islands are discrete regions of the genome because of two characteristics, their unusual CpG frequency and the fact that most commonly, although not always, they are free of methylation. Their high CpG content is generally attributed to the absence of methylation, the mechanism however that

keeps them free of methylation or the reason that this happens is not known. The observation that they are often found in housekeeping promoters seems to provide some biological meaning to their existence. The situation however becomes more complicated with the acknowledgment that CpG islands are not refractory to methylation, they are not confined to promoters and they are also found associated with tissue-specific genes.

The present study is the first high-throughput detailed analysis of CpG island methylation in mouse regardless of the location of the CpG islands in the genome. There are two main reasons that such analysis has not been conducted until now. Until very recently, the (almost) complete genome sequence of the mouse was not available. It is of course impossible to embark on any genome-wide CpG island scavenging without the genomic sequence. Additionally, an affinity purification technology of methylcytosines has not been available until recently. By taking advantage of the recent technical breakthroughs, the characteristics and methylation patterns of CpG islands in mouse brain were investigated and CpG islands that are potentially becoming methylated during *in vitro* differentiation of cells were identified.

Distribution of CpG islands in the genome

The number of CpG islands identified in this study is 20,755. This number is well below the 27,000 CpG islands estimated in the human genome (Mouse Genome Sequencing Consortium) and agrees with previous calculations that predicted CpG island depletion in mouse (Antequera and Bird 1993). However, the presently calculated number of mouse CpG islands is much higher than the estimate from the Mouse Genome Sequencing Consortium (2002), that reported 15,500 CpG islands.

The possibility that repetitive sequences could have contaminated the island population can be reasonably excluded as the genome had been repeat-masked before applying the CpG island algorithm. Moreover Ponger *et al.* (2001) had shown that contamination of the CpG island sequences with repetitive sequences is considerably reduced if, like in the present study, a window of 500 nucleotides is used.

It is possible that the present number is an overestimation because CpG islands were identified after *in silico* Mse I digestion of the genome. Mouse CpG islands are known to be less CpG and GC-rich which raises the possibility that Mse I

recognition sites (TTAA) are contained within them. This possibility was appreciated and the CpG islands were reassembled according to the distance and the overall length of the neighbouring Mse I fragments (section 4.2). It is possible that the reassembly had been incomplete. Calculation however of the GC content and the o/e ratio of the identified CpG islands provides support that the genomic regions had been correctly selected (Figure 4-2). Moreover, the agreement of some of the present results with other investigations as well the positive correlation of the CpG island content with the gene content in the chromosomes (Figure 4-3) supports the present estimation of the mouse CpG island number. Finally, some, partial, explanation to the discrepancy in the number of CpG islands could come from the fact that a later genome assembly used for the estimation of CpG islands in this study than the one that was used for the 15,500 estimate. As GC-regions are more difficult to sequence, this study might have included important CpG island genomic regions that were not available at the time of the previous calculation.

Half of the identified CpG islands are at the 5' region of genes, most of the remaining are equally divided in intragenic and intergenic regions and a small proportion lays at the 3' of genes. Previous analysis of the CpG islands in human chromosomes 21 and 22 (Takai and Jones 2002) showed that only17.4% of the CpG islands were located in 5' regions. The most important reason for the difference between the two studies appears to be a different approach in determining what consists 5' regions. In the present study known and predicted genes were included in the calculations, while Takai and Jones (2002) included only known genes. By doing so they identified an excess of 75% of CpG islands that could not be classified anywhere relatively to the transcription start site. In this study, the intergenic CpG islands were only 24.7% of the total. These numbers make it obvious that the inclusion of predicted genes in this study can account for the difference in 5'-associated CpG islands.

The number of genes (predicted and characterised) associated with the CpG islands that were identified is approximately 63% of the total (NCBI build 34). As predicted if the main functional role of CpG islands was transcriptional regulation, the majority of the CpG island-associated genes include them in their promoter regions (Table 4-1). This corresponds to approximately 52% of all the mouse genes,

which agrees with the number previously calculated (46.9%, Antequera and Bird 1993). A significant proportion of the CpG island-associated genes contains them in the gene body. About half of these genes also have a separate CpG island region identified in their 5'. Finally, a very small percentage of genes have a CpG island only at their 3'.

Sequence composition of methylated CpG islands in mouse

Previous studies have demonstrated that methylated CpG islands tend to be at the lower end of the o/e range (Fang *et al.* 2006; Bock *et al.* 2006; Weber *et al.* 2007). In the present study, the mean, median and shape of the distribution of the o/e values of the methylated CpG islands were compared with those from the entire set of CpG islands in order to identify any trends in the sequence composition of the methylated CpG islands in mouse. Statistical analysis produced contradictory results and no conclusion could be drawn regarding the CpG density of methylated CpG islands (section 4.5.1).

One explanation to the unclear CpG frequency character of the methylated CpG islands could come from the shape of their distribution (Figure 4-19). Due to skewing of the distribution, although there appears to be a local peak at lower o/e –and this is supported by the KS test– the bulk of the methylated CpG islands lays within higher CpG values. The results of the t-test and Wilcoxon test are supportive of the latter. It is almost as if there are two forces determining the CpG content of methylated CpG islands; one that causes the preferential methylation of CpG islands with a specific, lower than the bulk average CpG content and another that supports the methylation of a broader range of CpG islands of higher values.

As the purification of these CpG islands was based on their interaction with the MBD domain of MeCP2, which has been shown to have specificity for a single CpG (Klose *et al.* 2005), it is unlikely that the methodology had introduced bias in the selection of methylated CpG-islands. It is not impossible that in the mouse, methylation shows less preference for CpG-depleted CpG islands. Previous reports have shown that the mouse genome is more CpG-depleted than the human, presumably through mutation of methylcytosines (Matsuo *et al.* 1993). One can imagine that the process of mutation of methylated CpG islands may be close to equilibrium in this organism, in which the remaining CpG density of the CpG islands

is important for their function. If this is true, then it means that the function of CpG islands –whatever this might be– relies on a critical CpG density and could give clues for deciphering the role of CpG islands in the genome. Without further analysis however, it is impossible to draw any definite conclusions about the CpG density of methylated CpG islands in mouse from the present results.

Patterns of CpG island methylation in mouse brain

18.38% of all the identified CpG islands were found methylated in brain. This number compares very well to the report by Shiota *et al.* (2002) that calculated 16% of the CpG islands in mouse as being methylated in somatic tissues in comparison to stem and germ cells. Methylated CpG islands showed preference for genes that are involved in development, cell structure and communication. The importance of CpG island-associated genes in development has been implied before (Robinson *et al.* 2004) and the present data confirm this association.

Similarly to the results of Eckhardt *et al.* (2006), 15.63% of the 5' CpG islands were methylated in mouse brain. This corresponds to methylation of 34.72% of the genes with a 5' CpG island, or around 18% of all the genes in the genome (Table 4-4). The possibility that CpG island methylation is a specific transcriptional regulatory mechanism was explored further by analyzing the gene functions that were preferentially associated with methylated 5' CpG islands (Table 4-6). The genes that were identified to preferentially have a methylated CpG island at their 5' were typical of sensory receptors and are not expected to be important for neuronal function. Moreover, the neuron-specific genes that were identified to preferentially contain CpG islands (Table 4-2) were not enriched in the methylated CpG island fraction. It would be interesting to see if neuron-specific genes acquire methylation in non-neuronal tissues. If this is the case then these results are in accordance with a transcriptional role of CpG island methylation.

Although an important 23.32% of 3' CpG island-associated genes were methylated (Table 4-4), these genes were not enriched for any particular function. It is possible that CpG island methylation in these regions has no biological significance and only follows the methylation pattern of the rest of the genome. Alternatively, these CpG islands could be functionally related with some other gene.

Of course there is also a possibility that the role of these CpG islands is not related to the function of the gene but is more general.

Finally, it should be noted that intergenic CpG islands were methylated with almost the same frequency as 5' CpG islands (Table 4-4). It is possible, although not probable, that the robust gene prediction algorithms have missed approximately 5,000 genes that are present in what is today considered as intergenic regions. A more reasonable explanation however, would be that these regions only follow the methylation pattern of their surroundings. Spreading of DNA methylation has been suggested as a mechanism of methylation establishment in the genome (Turker 2002) and is supported by the phenomenon of position effect variegation in mice. Support for this assumption however, should come from experimental data. One way to test the assumption of DNA methylation spreading in intergenic regions would be to target foreign DNA with high CpG-density *e.g. Drosophila* genomic DNA, in intergenomic regions of mouse that are known to be euchromatic or heterochromatic. If the assumption is correct, then the CpG island-like sequence should acquire the methylation characteristics of its surroundings. Furthermore, direct methylation analysis of the non-CpG island genomic regions that are surrounding the intergenic CpG islands should be consistent with the methylation status of the latter, as determined in the present study.

Characterisation of *de novo* methylated CpG islands in differentiated ES cells

When the DNA from ES cells and embryoid bodies that was prepared with the MAP methodology, was hybridised to the CpG island microarrays, the results showed big variation between experiments. This prevented the genome-wide analysis of *de novo* methylated CpG islands. There are two possible reasons that might have caused variation between experiments. First, the ES cells had stayed in culture for different lengths of time before differentiation and this could have led to the appearance of aberrant methylation. Such a tendency for cells in culture to acquire DNA methylation has been described before (Antequera *et al.* 1990). However, the fact that these ES cells are routinely used for the production of transgenic ES cells in the laboratory testifies against such an explanation. Moreover, if different methylation patterns had been established during culturing of the ES cells, then the same patterns would be present in the embryoid bodies derived from them. However,

the observed variation between experiments had not improved when the pairs of ES cells and their differentiation derivatives were analysed together during normalisation of the data.

A more likely explanation for the experimental variation is that ES cells and embryoid bodies have very little methylation in comparison to adult tissues and the detected signal at the microarrays was mainly due to background hybridisation. Such a problem had been encountered before when the MAP methodology was applied for the purification of methylated CpG islands from sperm (R. Illingworth, personal communication). According to the model of methylation establishment during development (Shiota *et al.* 2002; Kremenskoy *et al.* 2003), ES cells that represent very early stages of development –as well as sperm– should have less methylation than adult tissues. If this is the case, then PCR amplification of the MAP-purified DNA before microarray hybridisation could be introducing noise to the data, making their interpretation difficult. Ways to avoid amplification of noise would be T7 RNA polymerase-mediated amplification (Liu *et al.* 2003) or using sufficient starting material so that amplification would not be required prior to labelling. Furthermore, verification of the microarray results by qPCR on DNA that has not been previously amplified would be a reliable test of whether the fragments are truly enriched after MAP. The MAP method itself has been shown to be reliable on purifying methylated DNA (Figure 4-10, Figure 4-14, Figure 4-15, Figure 4-18). All these taken together make it most probable that the majority of the CpG islands that were selected as being preferentially enriched in the embryoid bodies are indeed methylated (Figure 4-22, Table 4-8).

A transcription-driven mechanism for the determination of CpG island methylation could be acting during development

Out of the 5,883 genes in the mouse genome that fall under the umbrella of development-related, 2,771 have CpG islands (47.1%) (Table 4-2) and more than a quarter of these is methylated in brain (Table 4-5). On the other end, the other big class of CpG island-associated genes are genes with a maintenance or housekeeping function. According to the present analysis, this category alone makes up approximately 80% of all the CpG island-associated genes. As anticipated, these genes were not preferentially methylated. The observation that such a high

proportion of CpG island-associated genes has housekeeping functions, could be the explanation to why CpG islands were discovered through absence of methylation. The implication of this is that the CpG island DNA sequence could have no intrinsic tendency to be devoid of methylation. Instead, like in almost all examined promoters, it is the transcriptional status of the associated gene that determines the presence or absence of methylation at it.

An active role of transcriptional regulation on the methylation status of CpG islands can be seen by comparing housekeeping and developmental genes. These two main categories of CpG island-associated genes show opposite behaviours in brain regarding methylation; housekeeping genes seem to avoid methylation, while developmental genes preferentially acquire it (Table 4-5). This could be explained in terms of transcriptional regulation; both these classes of genes are of great importance for multicellular organisms, the housekeeping genes for cell survival and the developmental genes for organism survival. However, the demand for housekeeping genes is constant, while developmental genes become redundant after differentiation and body patterning has been completed. This could be determining their different fates regarding methylation.

Moreover, there was a group of CpG islands that appeared to become *de novo* methylated when pluripotent ES cells were induced to differentiate (Table 4-8). *In vitro* differentiation of ES cells is reported to mimic the events during early development and leads to the formation of the main cell lineages of the organism (Leahy *et al.* 1999; Choi *et al.* 2005 and present study, chapter 3). It can be reasonably assumed that these CpG islands are methylated in at least some of the *in vitro* derived cell lineages. Furthermore, identification of the methylation status of the same CpG islands in brain (Table 4-9), showed that –at least for the 5'-associated CpG islands– there was a direct correlation of the transcriptional status of the associated gene in brain with absence of methylation. These taken together, strengthen the relationship of CpG island methylation with absence of transcription and provide strong evidence for a developmental program for the establishment ofCpG island methylation.

The present study showed that CpG island methylation occurs throughout the mouse genome with similar frequencies, regardless of the position relative to the

transcription start site of a gene (Table 4-4). When the genes that are associated with intragenic CpG island methylation in brain were specifically analysed (Table 4-7), the gene ontology categories were very similar to the ones of the methylated 5' CpG island genes (receptor activity, signal transduction, sensory and chemosensory perception and olfaction). This can not be readily explained by concurrent methylation of the 5' and gene body, as analysis showed that, of all the genes that have CpG islands in both these regions, only approximately 10% show methylation in both (section 4.5.2). On the other hand, there was a distinct, although not absolute, correlation of the methylation status of the intergenic CpG islands in brain with the transcription status of their associated gene in the same tissue (Table 4-9). Moreover, CpG island methylation was established during *in vitro* differentiation of ES cells regardless of their location in the genome (Table 4-8). These data suggest that CpG island methylation in mouse is related with the transcription status of the associated genes regardless of their position relative to the associated gene.

Given the evidence from the present study, one can imagine a CpG island methylation process in which, as the initial totipotent zygote becomes more differentiated and the cell lineages committed, CpG island methylation serves as a special mechanism to irrevocably silence genes that are not anymore needed in the emerging cell lineage. This would lead to the methylation of only a group of the CpG islands that are associated with these genes in each tissue. Comparative methylation analyses of CpG islands in many adult tissues have shown that, indeed, although much of the methylation pattern is shared, there are also differences according to the cell type (Shiota *et al.* 2002; Kremenskoy *et al.* 2003; R. Illingworth, personal communication). Moreover, similar analysis of different regions and developmental stages of brain showed only small differences in CpG island methylation (1.7%) (Kawai *et al.* 1993).

The proposed scenario could be experimentally tested by examining the methylation status of CpG islands in many different adult cell types. These CpG islands could be the ones identified in this study as being *de novo* methylated in embryoid bodies (Table 4-8) as well as others that are associated with developmental genes. In a process analogous to the phylogenetic analysis of organisms, the most closely related cell types should have more similar methylation patterns than the ones

that diverged earlier in development. If the hypothesis is correct, then the tree produced by comparing the methylation status of CpG islands in the different tissues should be able to reproduce the known developmental relationships of the tested cell types.

Conclusions

The aim of this study was to provide a detailed account of the methylation status of CpG islands in mouse regardless of their position in the genome. In order to achieve this, the CpG islands in the mouse genome were identified by applying a novel CpG island prediction algorithm. It was found that half of the mouse CpG islands sit on the 5' prime end of genes, approximately 5% on the 3' prime end and approximately 23% are intragenic. The remaining 25% of the CpG islands are not associated with genes. The two main classes of the CpG island-associated genes were housekeeping and developmental, while two other interesting gene categories that were over-repressented in the CpG island-associated genes were olfactory and neuronal. Next, the methylated CpG islands from brain genomic DNA were purified by an affinity method and characterised. A total of 18.38% of the CpG islands were found to be methylated in brain the majority of which were located on 5' prime and intragenic regions. Developmental and olfactory genes were over-represented in the group of methylated CpG island-associated genes. Application of the same methodology showed that there is *de novo* methylation of CpG islands during *in vitro* differentiation of ES cells. Unfortunatelly, experimental limitations prohibited the global identification of the CpG islands that become methylated after the *in vitro* differentiation of ES cells. The experiment however provided a list of CpG islands that become methylated and it was shown that transcriptional inertia is most probably associated with them. Finally, all the data taken together, point to a developmental program of CpG island methylation establishment in mouse that is correlated with transcriptional silencing.

5. Discussion

The aim of this study has been to investigate how DNA methylation patterns are established during mouse development. *Oct4* has been used as a specific example of a developmentally regulated gene that becomes *de novo* methylated as it is transcriptionally repressed. The experiments described in the previous pages confirmed the previous finding that DNA methylation is dispensable and comes after the gene's initial downregulation. It has been further demonstrated that establishment of DNA methylation at the upstream region of *Oct4* starts at the proximal enhancer, which is dispensable for the gene's expression in ES cells (Yeom *et al.* 1996; Gu *et al.* 2005a), and is not targeted at the promoter as previously thought (Gidekel and Bergman 2002; Gu *et al.* 2006; Feldman *et al.* 2006). DNA methylation then spreads outwards, towards the distal enhancer and the promoter. Importantly, it has been shown that G9a does not function to recruit DNA methylation at the promoter of *Oct4*, as previously assumed (Feldman *et al.* 2006), but has a role in methylation spreading.

Perhaps the most important question that arises from these findings is whether the establishment of the DNA methylation pattern is active or passive. One experiment to test this would be to delete the proximal enhancer, from which DNA methylation is initiated. There are two main possible outcomes from such an experiment where the proximal enhancer of *Oct4* is deleted; methylation can be completely abolished from the upstream region of *Oct4*, or methylation will be established as normal, initiating from the new "middle region" between promoter and enhancer. The first case would be an indication that DNA methylation is targeted to the proximal enhancer, while the second would provide evidence of a more passive way for methylation establishment. Chromatin immunoprecipitation of the DNMT3s in the entire wild-type upstream region would strengthen the results; high abundance of DNMT3s at the proximal enhancer would be supporting the active methylation model, while more diffuse localisation would provide support for a passive methylation establishment.

In any case, elucidating the mechanism by which the DNA methylation pattern is established in the upstream region of *Oct4* consists an important finding. The only other case were the way that the DNA methylation pattern is established is known, is that of *Aprt* (Macleod *et al.* 1994; Mummaneni *et al.* 1998). At this gene,

transcription factor binding prevents methylation of the bound upstream region, presumably through steric hindrance of the DNMT3s. It is important that through *Oct4* and other systems of inducible DNA methylation we obtain more information on the molecular cascade that leads to the specific methylation patterns we observe in the genome. In this way we would perhaps be able to put forward a more generalised model of DNA methylation establishment, understand better the function of DNA methylation in the mammalian genome and, eventually, understand better the ay that the cell works.

Moreover, given evidence from this and other studies, a model was proposed to explain the observed *de novo* methylation pattern as well as the dispensability of DNA methylation for this gene's silencing. According to this model, DNA methylation acts as an anchor for transcription repressors that may be recruited through other mechanisms. Stabilisation of the orchestrated recruitment of repressors and chromatin remodelling complexes to the promoter leads to timely and efficient silencing. This model can explain why establishment of methylation at this gene would be important *in vivo*, during the coordinated differentiation of a multicellular organism, but dispensable for its transcriptional repression in cultured cells. However, as discussed in Chapter 3, this model needs to be experimentally tested in more detail.

For the investigation of genome-wide methylation patterns in mouse CpG islands, the CpG islands in the mouse genome were first identified by using a novel prediction algorithm. The results showed that half of the CpG islands are situated at the 5' end of genes, while the rest are almost equally divided between intergenic and intragenic regions. Only a very small proportion of CpG islands lies at the 3' of genes. Previous studies have indicated an association of CpG islands with housekeeping (Yamashita *et al.* 2005; Saxonov *et al.* 2006) and developmental genes (Robinson *et al.* 2004). Gene ontology analysis of the genes that are associated with the identified CpG islands in this study agreed with the previous reports, and also identified genes that are important for neuronal activity as being preferentially associated with CpG islands.

For the detection of methylated CpG islands, an affinity purification method was applied in conjunction with microarray hybridisation. Various control

experiments throughout the purification and hybridisation process have verified that this method is reliable in identifying methylated CpG islands. According to this analysis, approximately 43% of all the CpG island-associated genes are methylated in mouse brain. The bulk of these genes has developmental or tissue-specific functions. This is in accordance with a role of CpG island methylation in transcription, as well as the existence of a developmental program for CpG island methylation establishment.

The possibility that the CpG island methylation patterns are established during development was futher investigated by applying the same methodology on *in vitro* differentiated embryonic stem cells. Unfortunately, this approach failed to give information about the genome-wide pattern of CpG island methylation establishment. This was probably because the methylation levels achieved in the given time scale were very low and amplification of the material resulted in an increase of the noise-to-signal ratio. However, candidates for *de novo* methylated CpG islands were identified. When the methylation status of these CpG islands and the expression of their associated genes in brain was analysed, there was a very good correlation of the presence of methylation with transcriptional repression. Having established an *in vitro* model of early development where these CpG islands become *de novo* methylated, it will be very interesting to see whether and how methylation in these regions correlates temporally with transcription repression. If the mechanism of DNA methylation establishment in *Oct4* is widespread and acts on CpG islands, then it would be expected that methylation in these cases will follow the gene's repression. Moreover, this system provides a significant number of candidates for the study of the induced methylation of CpG islands. As outlined above, elucidation of the mechanism that leads to DNA methylation and the molecular events that surround it can provide invaluable knowledge on the function of the mammalian genome.

According to the detected methylation pattern of CpG islands in brain, almost equal proportions of the intergenic and 5'-associated CpG islands were found to be methylated. Taking into account the discovery that, in the case of *Oct4*, an element that lies more that 1 Kb upstream from the transcriptional start site –and not the immediate 5' region of the gene– preferentially becomes methylated, it is possible

that the intergenic methylated CpG islands have a role in the regulation of a gene from a distance. The possibility however, that these intergenic CpG islands only follow the methylation pattern of their surroundings cannot be excluded.

The evidence for a developmental program of DNA methylation establishment has accumulated in the recent years. The possibility that CpG islands and non-CpG island sequences acquire methylation through a similar mechanism can not be excluded. Moreover, it becomes clear that although DNA methylation is associated with transcriptional inertia, it does not appear to be indispensable for gene repression. Perhaps this is why certain organisms have managed to eliminated it from their genome. Further studies on specific examples of DNA methylation establishment, as well as genome-wide analyses of DNA methylation in different tissues and developmental stages, coupled with expression analysis of the associated genes, will shed more light on the impact that DNA methylation has in mammals.

Electronic resources

BiQ Analyser	http://biq-analyzer.bioinf.mpi-sb.mpg.de/
Bioconductor	http://biq-analyzer.bioinf.mpi-sb.mpg.de/
PANTHER, classification system	http://www.pantherdb.org/
Tree of life web project	http://www.tolweb.org/tree/
Symatlas	http://symatlas.gnf.org/SymAtlas/

Bibliography

"The Tree of Life Web Project."(1996-2006). from http://tolweb.org.

Aagaard L, Laible G, Selenko P, Schmid M, Dorn R, Schotta G, Kuhfittig S, Wolf A, Lebersorger A and Singh P B (1999). "Functional mammalian homologues of the Drosophila PEV-modifier Su(var)3.9 encode centromereassociated proteins which complex with the heterochromatin component M31." Embo Journal, 18(7): 1923.

Antequera F and Bird A (1993). "Number of CpG islands and genes in human and mouse." Proceedings- National Academy of Sciences USA, 90(24): 11995.

Antequera F and Bird A (1999). "CpG islands as genomic footprints of promoters that are associated with replication origins." Current Biology, 9(17): R661-R667.

Antequera F, Boyes J and Bird A (1990). "High levels of De Novo methylation and altered chromatin structure at CpG islands in cell lines." Cell, 62: 503-514.

Aoki A, Suetake I, Miyagawa J, Fujio T, Chijiwa T, Sasaki H and Tajima S (2001). "Enzymatic properties of de novo-type mouse DNA (cytosine-5) methyltransferases." Nucleic Acids Research, 29(17): 3506-3512.

Aoto T, Saitoh N, Ichimura T, Niwa H and Nakao M (2006). "Nuclear and chromatin reorganization in the MHC-Oct3/4 locus at developmental phases of embryonic stem cell differentiation." Developmental Biology, 298(2): 354-367.

Arabidopsis Genome Initiative, (2000). "Analysis of the genome sequence of the flowering plant Arabidopsis thaliana" Nature, 408: 796-815.

Bacher C P, Guggiari M, Brors B, Augui S, Clerc P, Avner P, Eils R and Heard E (2006). "Transient colocalization of X-inactivation centres accompanies the initiation of X inactivation." Nature Cell Biology, 8(3): 293-299.

Bachman K E, Rountree M R and Baylin S B (2001). "Dnmt3a and Dnmt3b are transcriptional repressors that exhibit unique localization properties to heterochromatin." Journal of Biological Chemistry, 276(34): 32282-32287.

Bai S, Ghoshal K, Datta J, Majumder S, Yoon S O and Jacob S T (2005). "DNA Methyltransferase 3b Regulates Nerve Growth Factor-Induced Differentiation of PC12 Cells by Recruiting Histone Deacetylase 2." Molecular and Cellular Biology, 25: 751-766.

Ball T J, Gross D S and Garrard W T (1983). "5-Methylcytosine is localised in nucleosomes that contain histone H1." Proceedings- National Academy of Sciences USA, 80: 5490-5494.
Barnea E and Bergman Y (2000). "Synergy of SF1 and RAR in activation of Oct-3/4 promoter." Journal of Biological Chemistry, 275(9): 6608-6619.
Barreto G, Schafer A, Marhold J, Stach D, Swaminathan S K, Handa V, Doderlein G, Maltry N, Wu W, Lyko F and Niehrs C (2007). "Gadd45a promotes epigenetic gene activation by repair-mediated DNA demethylation." Nature, 445(7128): 671-675.
Beaujean N, Hartshorne G, Cavilla J, Taylor J, Gardner J, Wilmut I, Meehan R and Young L (2004). "Non-conservation of mammalian preimplantation methylation dynamics." Current Biology, 14(7): R266-R267.
Bell A C and Felsenfeld G (2000). "Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene." Nature(6785): 482-485.
Bell A C, West A G and Felsenfeld G (1999). "The Protein CTCF Is Required for the Enhancer Blocking Activity of Vertebrate Insulators." Cell, 98(3): 387-396.
Ben-Shushan E, Thompson J R, Gudas L J and Bergman Y (1998). "Rex-1, a Gene Encoding a Transcription Factor Expressed in the Early Embryo, Is Regulated via Oct-3/4 and Oct-6 Binding to an Octamer Site and a Novel Protein, Rox-1, Binding to an Adjacent Site." Molecular and Cellular Biology, 18(4): 1866-1878.
Berger J, Sansom O, Clarke A and Bird A (2007). "MBD2 Is Required for Correct Spatial Gene Expression in the Gut." Molecular and Cellular Biology, 27(11): 4049-4057.
Bernstein B E, Humphrey E L, Erlich R L, Schneider R, Bouman P, Liu J S, Kouzarides T and Schreiber S L (2002). "Methylation of histone H3 Lys 4 in coding regions of active genes." Proceedings- National Academy of Sciences USA, 99(13): 8695-8700.
Bernstein B E, Kamal M, Lindblad-Toh K, Bekiranov S, Bailey D K, Huebert D J, McMahon S, Karlsson E K, Kulbokas Iii E J and Gingeras T R (2005). "Genomic Maps and Comparative Analysis of Histone Modifications in Human and Mouse." Cell, 120(2): 169-181.
Bestor T H (1990). "DNA methylation:evolution of a bacterial immune function into a regulator of gene expression and genome structure in higher eukaryotes." Philosophical Transactions of the Royal Society of London, B Biological Sciences, 326: 179-187.
Bird A, Taggart M, Frommer M, Miller O J and Macleod D (1985). "A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA." Cell, 40(1): 91-99.
Bird A, Taggart M and Macleod D (1981). "Loss of rDNA methylation accompanies the onset of ribosomal gene activity in early development of X. laevis." Cell, 26: 381-390.
Bird A P (1980). "DNA methylation and the frequency of CpG in animal DNA." Nucl. Acids Res., 8(7): 1499-1504.
Bird A P (1986). "CpG-rich islands and the function of DNA methylation." Nature, 321(6067): 209-213.
Bird A P (1995). "Gene number, noise reduction and biological complexity." Trends in Genetics, 11(3): 94.

Birke M, Schreiner S, Garcia-Cuellar M-P, Mahr K, Titgemeyer F and Slany R K (2002). "The MT domain of the proto-oncoprotein MLL binds to CpG-containing DNA and discriminates against methylation." Nucleic Acids Research, 30(4): 958-965.

Bock C, Paulsen M, Tierling S, Mikeska T, Lengauer T, Walter J and rn (2006). "CpG Island Methylation in Human Lymphocytes Is Highly Correlated with DNA Sequence, Repeats, and Predicted DNA Structure." PLoS Genetics, 2(3): e26.

Boyer L A, Lee T I, Cole M F, Johnstone S E, Levine S S, Zucker J P, Guenther M G, Kumar R M, Murray H L and Jenner R G (2005). "Core Transcriptional Regulatory Circuitry in Human Embryonic Stem Cells." Cell, 122(6): 947-956.

Brero A, Easwaran H P, Nowak D, Grunewald I, Cremer T, Leonhardt H and Cardoso M C (2005). "Methyl CpG-binding proteins induce large-scale chromatin reorganization during terminal differentiation." Journal of Cell Biology, 169(5): 733-743.

Campanero M R, Armstrong M I and Flemington E K (2000). "CpG methylation as a mechanism for the regulation of E2F activity." Proceedings- National Academy of Sciences USA, 97(12): 6481-6486.

Campoy F J, Meehan R R, McKay S, Nixon J and Bird A (1995). "Binding of Histone H1 to DNA Is Indifferent to Methylation at CpG Sequences." Journal of Biological Chemistry, 270(44): 26473.

Cao X, Aufsatz W, Zilberman D, Mette M F, Huang M S, Matzke M and Jacobsen S E (2003). "Role of the DRM and CMT3 Methyltransferases in RNA-Directed DNA Methylation." Current Biology, 13(24): 2212-2217.

Cardoso M C and Leonhardt H (1999). "DNA Methyltransferase Is Actively Retained in the Cytoplasm during Early Development." Journal of Cell Biology, 147(1): 25-32.

Carlone D L, Lee J H, Young S R L, Dobrota E, Butler J S, Ruiz J and Skalnik D G (2005). "Reduced Genomic Cytosine Methylation and Defective Cellular Differentiation in Embryonic Stem Cells Lacking CpG Binding Protein." Molecular and Cellular Biology, 25: 4881-4891.

Carlone D L and Skalnik D G (2001). "CpG Binding Protein Is Crucial for Early Embryonic Development." Molecular and Cellular Biology, 21(22): 7601-7606.

Carlson L L, Page A W and Bestor T H (1992). "Properties and localization of DNA methyltransferase in preimplantation mouse embryos: implications for genomic imprinting." Genes and Development, 6(12b): 2536-2541.

Catena R, Tiveron C, Ronchi A, Porta S, Ferri A, Tatangelo L, Cavallaro M, Favaro R, Ottolenghi S and Reinbold R (2004). "Conserved POU Binding DNA Sites in the Sox2 Upstream Enhancer Regulate Gene Expression in Embryonic and Neural Stem Cells." Journal of Biological Chemistry, 279: 41846-41857.

Chambers I, Colby D, Robertson M, Nichols J, Lee S, Tweedie S and Smith A (2003). "Functional Expression Cloning of Nanog, a Pluripotency Sustaining Factor in Embryonic Stem Cells." Cell, 113(5): 643-656.

Chen T, Ueda Y, Dodge J E, Wang Z and Li E (2003). "Establishment and Maintenance of Genomic Methylation Patterns in Mouse Embryonic Stem

Cells by Dnmt3a and Dnmt3b." Molecular and Cellular Biology, 23(16): 5594-5605.
Chen T, Ueda Y, Xie S and Li E (2002). "A Novel Dnmt3a Isoform Produced from an Alternative Promoter Localizes to Euchromatin and Its Expression Correlates with Active de Novo Methylation." Journal of Biological Chemistry, 277(41): 38746-38754.
Choi D, Lee H J, Jee S, Jin S, Koo S K, Paik S S, Jung S C, Hwang S Y, Lee K S and Oh B (2005). "In Vitro Differentiation of Mouse Embryonic Stem Cells: Enrichment of Endodermal Cells in the Embryoid Body." Stem Cells, 23: 817-827.
Chu M W, Siegmund K D, Eckstam C L, Kim J Y, Yang A S, Kanel G C, Tavaré S and Shibata D (2007). "Lack of increases in methylation at three CpG-rich genomic loci in non-mitotic adult tissues during aging." BMC medical genetics, 8(50).
Clark S J, Harrison J and Frommer M (1995). "CpNpG methylation in mammalian cells." Nature Genetics, 10(1): 20.
Cooper D N, Taggart M H and Bird A P (1983). "Unmethlated domains in vertebrate DNA." Nucleic Acids Research, 11(3): 647-658.
Craig J M and Bickmore W A (1994). "The distribution of CpG islands in mammalian chromosomes." Nature Genetics, 7(3): 376-382.
Craig J M, Earle E, Canham P, Wong L H, Anderson M and Choo K H A (2003). "Analysis of mammalian proteins involved in chromatin modification reveals new metaphase centromeric proteins and distinct chromosomal distribution patterns." Human Molecular Genetics, 12(23): 3109-3121.
Cross S H, Charlton J A, Nan X and Bird A P (1994). "Purification of CpG islands using a methylated DNA binding column." Nat Genet, 6(3): 236-244.
Cui X S, Shen X H and Kim N H (2007). "Dicer1 expression in preimplantation mouse embryos: Involvement of Oct3/4 transcription at the blastocyst stage." Biochemical and Biophysical Research Communications, 352(1): 231-236.
Czermin B, Melfi R, McCabe D, Seitz V, Imhof A and Pirrotta V (2002). "Drosophila Enhancer of Zeste/ESC Complexes Have a Histone H3 Methyltransferase Activity that Marks Chromosomal Polycomb Sites." Cell, 111(2): 185-196.
Daniel J M and Reynolds A B (1999). "The Catenin p120ctn Interacts with Kaiso, a Novel BTB/POZ Domain Zinc Finger Transcription Factor." Molecular and Cellular Biology, 19(5): 3614-3623.
Davey C, Fraser R, Smolle M, Simmen M W and Allan J (2003). "Nucleosome Positioning Signals in the DNA Sequence of the Human and Mouse H19 Imprinting Control Regions." Journal of Molecular Biology, 325(5): 873-887.
Davey C, Pennings S and Allan J (1997). "CpG Methylation Remodels Chromatin Structure in vitro." Journal of Molecular Biology, 267(2): 276-288.
Davey C S, Pennings S, Reilly C, Meehan R R and Allan J (2004). "A determining influence for CpG dinucleotides on nucleosome positioning in vitro." Nucleic Acids Research, 32(14): 4322-4331.
Davis T L, Yang G J, McCarrey J R and Bartolomei M S (2000). "The H19 methylation imprint is erased and re-established differentially on the parental alleles during male germ cell development." Human Molecular Genetics, 9(19): 2885-2894.

Dean W, Santos F, Stojkovic M, Zakhartchenko V, Walter J, Wolf E and Reik W (2001). "Conservation of methylation reprogramming in mammalian development: Aberrant reprogramming in cloned embryos." Proceedings-National Academy of Sciences USA, 98(24): 13734-13738.

Delgado S, Gomez M, Bird A and Antequera F (1998). "Initiation of DNA replication at CpG islands in mammalian chromosomes." Embo Journal, 17(8): 2426-2435.

Di Croce L, Raker V A, Corsaro M, Fazi F, Fanelli M, Faretta M, Fuks F, Coco F L, Kouzarides T, Nervi C, Minucci S and Pelicci P G (2002). "Methyltransferase Recruitment and DNA Hypermethylation of Target Promoters by an Oncogenic Transcription Factor." Science, 295(5557): 1079-1082.

Dillon S, Zhang X, Trievel R and Cheng X (2005). "The SET-domain protein superfamily: protein lysine methyltransferases." Genome Biology, 6(8): 227.

Dodge J E, Ramsahoye B H, Wo Z G, Okano M and Li E (2002). "De novo methylation of MMLV provirus in embryonic stem cells: CpG versus non-CpG methylation." Gene, 289(1-2): 41-48.

Douet V, Heller M B and Le Saux O (2007). "DNA methylation and Sp1 binding determine the tissue-specific transcriptional activity of the mouse Abcc6 promoter." Biochemical and Biophysical Research Communications, 354(1): 66-71.

Duncan B K and Miller J H (1980). "Mutagenic deamination of cytosine residues in DNA." Nature, 287(5782): 560-561.

Duret L and Galtier N (2000). "The Covariation Between TpA Deficiency, CpG Deficiency, and G+C Content of Human Isochores Is Due to a Mathematical Artifact." Molecular Biology and Evolution, 17(11): 1620-1625.

Eckhardt F, Lewin J, Cortese R, Rakyan V K, Attwood J, Burger M, Burton J, Cox T V, Davies R and Down T A (2006). "DNA methylation profiling of human chromosomes 6, 20 and 22." Nature Genetics, 38(12): 1378-1385.

Edwards Y H (1990). "CpG islands in genes showing tissue-specific expression." Philosophical Transactions of the Royal Society of London, B Biological Sciences, 326: 207-215.

Engel N, Thorvaldsen J L and Bartolomei M S (2006). "CTCF binding sites promote transcription initiation and prevent DNA methylation on the maternal allele at the imprinted H19/Igf2 locus." Human Molecular Genetics, 15: 2945-2954.

Fang F, Fan S, Zhang X and Zhang M Q (2006). "Predicting methylation status of CpG islands in the human brain." Bioinformatics, 22: 2204-2209.

Fatemi M, Hermann A, Pradhan S and Jeltsch A (2001). "The Activity of the Murine DNA Methyltransferase Dnmt1 is Controlled by Interaction of the Catalytic Domain with the N-terminal Part of the Enzyme Leading to an Allosteric Activation of the Enzyme after Binding to Methylated DNA." Journal of Molecular Biology, 309(5): 1189-1199.

Feldman N, Gerson A, Fang J, Li E, Zhang Y, Shinkai Y, Cedar H and Bergman Y (2006). "G9a-mediated irreversible epigenetic inactivation of Oct-3/4 during early embryogenesis." Nature Cell Biology, 8(2): 188-194.

Feng Q and Zhang Y (2001). "The MeCP1 complex represses transcription through preferential binding, remodeling, and deacetylating methylated nucleosomes." Genes and Development, 15(7): 827-832.

Field L M (2000). "Methylation and expression of amplified esterase genes in the aphid Myzus persicae (Sulzer)." Biochemical Journal, 349(3): 863-868.
Filion G J P, Zhenilo S, Salozhin S, Yamada D, Prokhortchouk E and Defossez P-A (2006). "A Family of Human Zinc Finger Proteins That Bind Methylated DNA and Repress Transcription." Molecular and Cellular Biology, 26(1): 169-181.
Fraga M F, Ballestar E, Paz M F, Ropero S, Setien F, Ballestar M L, Heine-Suner D, Cigudosa J C, Urioste M and Benitez J (2005). "Epigenetic differences arise during the lifetime of monozygotic twins." Proceedings- National Academy of Sciences USA, 102: 10604-10609.
Freitag M, Williams R L, Kothe G O and Selker E U (2002). "A cytosine methyltransferase homologue is essential for repeat-induced point mutation in Neurospora crassa." Proceedings- National Academy of Sciences USA, 99(13): 8802-8807.
Frommer M, McDonald L E, Millar D S, Collis C M, Watt F, Grigg G W, Molloy P L and Paul C L (1992). "A Genomic Sequencing Protocol that Yields a Positive Display of 5-Methylcytosine Residues in Individual DNA Strands." Proceedings of the National Academy of Science USA, 89(5): 1827-1831.
Fuhrmann G, Chung A C K, Jackson K J, Hummelke G, Baniahmad A, Sutter J, Sylvester I, Scholer H R and Cooney A J (2001). "Mouse Germline Restriction of Oct4 Expression by Germ Cell Nuclear Factor." Developmental Cell, 1(3): 377-387.
Fujita N, Takebayashi S I, Okumura K, Kudo S, Chiba T, Saya H and Nakao M (1999). "Methylation-Mediated Transcriptional Silencing in Euchromatin by Methyl-CpG Binding Protein MBD1 Isoforms." Molecular and Cellular Biology, 19(9): 6415-6426.
Fujita N, Watanabe S, Ichimura T, Tsuruzoe S, Shinkai Y, Tachibana M, Chiba T and Nakao M (2003). "Methyl-CpG Binding Domain 1 (MBD1) Interacts with the Suv39h1-HP1 Heterochromatic Complex for DNA Methylation-based Transcriptional Repression." Journal Biological Chemistry, 278(26): 24132-24138.
Fuks F, Burgers W A, Brehm A, Hughes-Davies L and Kouzarides T (2000). "DNA methyltransferase Dnmt1 associates with histone deacetylase activity." Nature Genetics, 24(1): 88-91.
Fuks F, Burgers W A, Godin N, Kasai M and Kouzarides T (2001). "Dnmt3a binds deacetylases and is recruited by a sequence-specific repressor to silence transcription." Embo Journal, 20(10): 2536-2544.
Fuks F, Hurd P J, Deplus R and Kouzarides T (2003a). "The DNA methyltransferases associate with HP1 and the SUV39H1 histone methyltransferase." Nucleic Acids Research, 31(9): 2305-2312.
Fuks F, Hurd P J, Wolf D, Nan X, Bird A P and Kouzarides T (2003b). "The Methyl-CpG-binding Protein MeCP2 Links DNA Methylation to Histone Methylation." Journal Biological Chemistry, 278(6): 4035-4040.
Gallais R, Demay F, Barath P, Finot L, Jurkowska R, Le Guevel R, Gay F, Jeltsch A, Metivier R and Salbert G (2007). "Dnmt 3a and 3b associate with the nuclear orphan receptor COUP-TFI during gene activation." Molecular Endocrinology: me.2006-0490.

Gardiner-Garden M and Frommer M (1987). "CpG Islands in vertebrate genomes." Journal of Molecular Biology, 196(2): 261-282.
Gehring M, Huh J H, Hsieh T-F, Penterman J, Choi Y, Harada J J, Goldberg R B and Fischer R L (2006). "DEMETER DNA Glycosylase Establishes MEDEA Polycomb Gene Self-Imprinting by Allele-Specific Demethylation." Cell, 124(3): 495-506.
Gidekel S and Bergman Y (2002). "A Unique Developmental Pattern of Oct-3/4 DNA Methylation Is Controlled by a cis-Demodification Element." Journal of Biological Chemistry, 277(37): 34521-34530.
Gilbert N, Lutz-Prigge S and Moran J V (2002). "Genomic Deletions Created upon LINE-1 Retrotransposition." Cell, 110(3): 315-325.
Gilbert N, Thomson I, Boyle S, Allan J, Ramsahoye B and Bickmore W A (2007). "DNA methylation affects nuclear organization, histone modifications, and linker histone binding but not chromatin compaction." Journal of Cell Biology, 177: 401-412.
Ginsburg M, Snow M H and McLaren A (1990). "Primordial germ cells in the mouse embryo during gastrulation." Development, 110(2): 521-528.
Goll M G and Bestor T H (2005). "Eukaryotic cytosine methyltransferases." Annual Review of Biochemistry, 74(1): 481-514.
Goll M G, Kirpekar F, Maggert K A, Yoder J A, Hsieh C-L, Zhang X, Golic K G, Jacobsen S E and Bestor T H (2006). "Methylation of tRNAAsp by the DNA Methyltransferase Homolog Dnmt2." Science, 311(5759): 395-398.
Gomez M and Antequera F (1999). "Organization of DNA replication origins in the fission yeast genome." Embo Journal, 18(20): 5683-5690.
Goncalves I, Duret L and Mouchiroud D (2000). "Nature and Structure of Human Genes that Generate Retropseudogenes." Genome Research, 10(5): 672-678.
Gowher H and Jeltsch A (2002). "Molecular Enzymology of the Catalytic Domains of the Dnmt3a and Dnmt3b DNA Methyltransferases." J. Biol. Chem., 277(23): 20409-20414.
Grunau C, Hindermann W and Rosenthal A (2000). "Large-scale methylation analysis of human genomic DNA reveals tissue-specific differences between the methylation profiles of genes and pseudogenes." Human Molecular Genetics, 9(18): 2651-2664.
Gu P, Goodwin B, Chung A C K, Xu X, Wheeler D A, Price R R, Galardi C, Peng L, Latour A M and Koller B H (2005a). "Orphan Nuclear Receptor LRH-1 Is Required To Maintain Oct4 Expression at the Epiblast Stage of Embryonic Development." Molecular and Cellular Biology, 25: 3492-3505.
Gu P, Le Menuet D, Chung A C K and Cooney A J (2006). "Differential Recruitment of Methylated CpG Binding Domains by the Orphan Receptor GCNF Initiates the Repression and Silencing of Oct4 Expression." Molecular and Cellular Biology, 26: 9471-9483.
Gu P, LeMenuet D, Chung A C K, Mancini M, Wheeler D A and Cooney A J (2005b). "Orphan Nuclear Receptor GCNF Is Required for the Repression of Pluripotency Genes during Retinoic Acid-Induced Embryonic Stem Cell Differentiation." Molecular and Cellular Biology, 25: 8507-8519.
Guanchao Jiang F Y C S M E (2004). "Histone modification in constitutive heterochromatin versus unexpressed euchromatin in human cells." Journal of Cellular Biochemistry, 93(2): 286-300.

Guy J, Hendrich B, Holmes M, Martin J E and Bird A (2001). "A mouse Mecp2-null mutation causes neurological symptoms that mimic Rett syndrome." Nature Genetics, 27(3): 322-326.
Haines T R, Rodenhiser D I and Ainsworth P J (2001). "Allele-Specific Non-CpG Methylation of the Nf1 Gene during Early Mouse Development." Developmental Biology, 240(2): 585-598.
Hajkova P, Erhardt S, Lane N, Haaf T, El-Maarri O, Reik W, Walter J and Surani M A (2002). "Epigenetic reprogramming in mouse primordial germ cells." Mechanisms of Development, 117(1-2): 15-23.
Handyside A H, O'Neill G T, Jones M and Hooper M L (1989). "Use of BRL-conditioned medium in combination with feeder cells to isolate a diploid embryonal stem cell line." Roux's Archives of Developmental Biology, 198: 48-55.
Hansen R S, Wijmenga C, Luo P, Stanek A M, Canfield T K, Weemaes C M R and Gartler S M (1999). "The DNMT3B DNA methyltransferase gene is mutated in the ICF immunodeficiency syndrome." Proceedings- National Academy of Sciences USA, 96(25): 14412-14417.
Hark A T, Schoenherr C J, Katz D J, Ingram R S, Levorse J M and Tilghman S M (2000). "CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus." Nature(6785): 486-489.
Hashimshony T, Zhang J, Keshet I, Bustin M and Cedar H (2003). "The role of DNA methylation in setting up chromatin structure during development." Nature Genetics, 34(2): 187-192.
Hata K, Okano M, Lei H and Li E (2002). "Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice." Development, 129(8): 1983-1993.
Hattori N, Nishino K, Ko Y g, Ohgane J, Tanaka S and Shiota K (2004). "Epigenetic Control of Mouse Oct-4 Gene Expression in Embryonic Stem Cells and Trophoblast Stem Cells." Journal of Biological Chemistry, 279(17): 17063-17069.
Hellman A and Chess A (2007). "Gene Body-Specific Methylation on the Active X Chromosome." Science, 315(5815): 1141-1142.
Hendrich B and Bird A (1998). "Identification and Characterization of a Family of Mammalian Methyl-CpG Binding Proteins." Molecular and Cellular Biology, 18(11): 6538-6547.
Hendrich B, Guy J, Ramsahoye B, Wilson V A and Bird A (2001). "Closely related proteins MBD2 and MBD3 play distinctive but interacting roles in mouse development." Genes and Development, 15(6): 710-723.
Hendrich B, Hardeland U, Ng H-H, Jiricny J and Bird A (1999). "The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites." Nature, 401(6750): 301-304.
Hermann A, Goyal R and Jeltsch A (2004). "The Dnmt1 DNA-(cytosine-C5)-methyltransferase Methylates DNA Processively with High Preference for Hemimethylated Target Sites." Journal of Biological Chemistry, 279: 48350-48359.
Hogan B, Beddington R, Constantini F and Lacy E (1994). Manipulating the Mouse Embryo. Cold Spring Harbor, Cold Spring Harbor Press.

Hopfner R, Mousli M, Jeltsch J-M, Voulgaris A, Lutz Y, Marin C, Bellocq J-P, Oudet P and Bronner C (2000). "ICBP90, a Novel Human CCAAT Binding Protein, Involved in the Regulation of Topoisomerase II{{alpha}} Expression." Cancer Research, 60(1): 121-128.

Horike S-i, Cai S, Miyano M, Cheng J-F and Kohwi-Shigematsu T (2005). "Loss of silent-chromatin looping and impaired imprinting of DLX5 in Rett syndrome." Nature Genetics, 37(1): 31-40.

Howell C Y, Bestor T H, Ding F, Latham K E, Mertineit C, Trasler J M and Chaillet J R (2001). "Genomic Imprinting Disrupted by a Maternal Effect Mutation in the Dnmt1 Gene." Cell, 104(6): 829-838.

Human Genome Sequencing International Consortium, (2004). "Finishing the euchromatic sequence of the human genome" Nature, 431: 931-945.

Humpherys D, Eggan K, Akutsu H, Hochedlinger K, Rideout W M, Biniszkiewicz D, Yanagimachi R and Jaenisch R (2001). "Epigenetic Instability in ES Cells and Cloned Mice." Science(5527): 95-97.

Jabbari K and Bernardi G (2004). "Cytosine methylation and CpG, TpG (CpA) and TpA frequencies." Gene, 333: 143-149.

Jabbari K, Caccio S, Pais de Barros J P, Desgres J and Bernardi G (1997). "Evolutionary changes in CpG and methylation levels in the genome of vertebrates." Gene, 205(1-2): 109-118.

Jablonka E and Lamb M J (1995). Epigenetic inheritance and evolution. Oxford, Oxford University Press.

Jackson M, Krassowska A, Gilbert N, Chevassut T, Forrester L, Ansell J and Ramsahoye B (2004). "Severe Global DNA Hypomethylation Blocks Differentiation and Induces Histone Hyperacetylation in Embryonic Stem Cells." Molecular and Cellular Biology, 24(20): 8862-8871.

Jareborg N, Birney E and Durbin R (1999). "Comparative Analysis of Noncoding Regions of 77 Orthologous Mouse and Human Gene Pairs." Genome Research, 9(9): 815-824.

Jia D, Jurkowska R Z, Zhang X, Jeltsch A and Cheng X (2007). "Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation." Nature, advanced online publication.

Jiang C, Han L, Su B, Li W-H and Zhao Z (2007). "Features and Trend of Loss of Promoter-Associated CpG Islands in the Human and Mouse Genomes." Molecular Biology and Evolution, 24(9): 1991-2000.

Jin S-G, Jiang C-L, Rauch T, Li H and Pfeifer G P (2005). "MBD3L2 Interacts with MBD3 and Components of the NuRD Complex and Can Oppose MBD2-MeCP1-mediated Methylation Silencing." Journal Biological Chemistry, 280(13): 12700-12709.

Johnson L M, Bostick M, Zhang X, Kraft E, Henderson I, Callis J and Jacobsen S E (2007). "The SRA Methyl-Cytosine-Binding Domain Links DNA and Histone Methylation." Current Biology, 17(4): 379-384.

Jones P L, Jan Veenstra G C, Wade P A, Vermaak D, Kass S U, Landsberger N, Strouboulis J and Wolffe A P (1998). "Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription." Nature Genetics, 19(2): 187-191.

Jørgensen H F, Ben-Porath I and Bird1 A P (2004). "Mbd1 Is Recruited to both Methylated and Nonmethylated CpGs via Distinct DNA Binding Domains." Molecular and Cellular Biology, 24(8): 3387–3395.
Joshi A A and Struhl K (2005). "Eaf3 Chromodomain Interaction with Methylated H3-K36 Links Histone Deacetylation to Pol II Elongation." Molecular Cell, 20(6): 971-978.
Kaji K, Nichols J and Hendrich B (2007). "Mbd3, a component of the NuRD co-repressor complex, is required for development of pluripotent cells." Development, 134(6): 1123-1132.
Kanellopoulou C, Muljo S A, Kung A L, Ganesan S, Drapkin R, Jenuwein T, Livingston D M and Rajewsky K (2005). "Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing." Genes and Development, 19(4): 489-501.
Kang M I, Rhyu M G, Kim Y H, Jung Y C, Hong S J, Cho C S and Kim H S (2006). "The length of CpG islands is associated with the distribution of Alu and L1 retroelements." Genomics, 87(5): 580-590.
Kareta M S, Botello Z M, Ennis J J, Chou C and Chedin F (2006). "Reconstitution and Mechanism of the Stimulation of de Novo Methylation by Human DNMT3L." Journal of Biological Chemistry, 281: 25893.
Kass S U, Landsberger N and Wolffe A P (1997). "DNA methylation directs a time-dependent repression of transcription initiation." Current Biology, 7(3): 157-165.
Kato M, Miura A, Bender J, Jacobsen S E and Kakutani T (2003). "Role of CG and Non-CG Methylation in Immobilization of Transposons in Arabidopsis." Current Biology, 13(5): 421-426.
Kawai J, Hirotsune S, Hirose K and Fushiki S (1993). "Methylation profiles of genomic DNA of mouse developmental brain detected by restriction landmark genomic scanning (RLGS) method." Nucleic Acids Research, 21(24): 5604.
Kim G D, Ni J, Kelesoglu N, Roberts R J and Pradhan S (2002). "Co-operation and communication between the human maintenance and de novo DNA (cytosine-5) methyltransferases." Embo Journal, 21(15): 4183-4195.
Kim S W, Park J-I, Spring C M, Sater A K, Ji H, Otchere A A, Daniel J M and McCrea P D (2004). "Non-canonical Wnt signals are modulated by the Kaiso transcriptional repressor and p120-catenin." Nature Cell Biology, 6(12): 1212-1220.
Kimura H and Shiota K (2003). "Methyl-CpG-binding Protein, MeCP2, Is a Target Molecule for Maintenance DNA Methyltransferase, Dnmt1." Journal Biological Chemistry, 278(7): 4806-4812.
Klose R J and Bird A P (2004). "MeCP2 Behaves as an Elongated Monomer That Does Not Stably Associate with the Sin3a Chromatin Remodeling Complex." Journal Biological Chemistry, 279(45): 46490-46496.
Klose R J, Sarraf S A, Schmiedeberg L, McDermott S M, Stancheva I and Bird A P (2005). "DNA Binding Selectivity of MeCP2 Due to a Requirement for A/T Sequences Adjacent to Methyl-CpG." Molecular Cell, 19(5): 667-678.
Kondo E, Gu Z, Horii A and Fukushige S (2005). "The Thymine DNA Glycosylase MBD4 Represses Transcription and Is Associated with Methylated

p16^I^N^K^4^a and hMLH1 Genes." Molecular and Cellular Biology, 25: 4388-4396.
Kremenskoy M, Kremenska Y, Ohgane J, Hattori N, Tanaka S, Hashizume K and Shiota K (2003). "Genome-wide analysis of DNA methylation status of CpG islands in embryoid bodies, teratomas, and fetuses." Biochemical and Biophysical Research Communications, 311(4): 884-890.
Kriaucionis S and Bird A (2004). "The major form of MeCP2 has a novel N-terminus generated by alternative splicing." Nucleic Acids Research, 32(5): 1818-1823.
Kriaucionis S, Paterson A, Curtis J, Guy J, MacLeod N and Bird A (2006). "Gene Expression Analysis Exposes Mitochondrial Abnormalities in a Mouse Model of Rett Syndrome." Molecular and Cellular Biology, 26(13): 5033-5042.
Krogan N J, Kim M, Tong A, Golshani A, Cagney G, Canadien V, Richards D P, Beattie B K, Emili A, Boone C, Shilatifard A, Buratowski S and Greenblatt J (2003). "Methylation of Histone H3 by Set2 in Saccharomyces cerevisiae Is Linked to Transcriptional Elongation by RNA Polymerase II." Molecular and Cellular Biology, 23(12): 4207-4218.
Kunert N, Marhold J, Stanke J, Stach D and Lyko F (2003). "A Dnmt2-like protein mediates DNA methylation in Drosophila." Development, 130(21): 5083-5090.
La Salle S, Mertineit C, Taketo T, Moens P B, Bestor T H and Trasler J M (2004). "Windows for sex-specific methylation marked by DNA methyltransferase expression profiles in mouse germ cells." Developmental Biology, 268(2): 403-415.
Labhart P (1994). "Negative and positive effects of CpG-methylation on Xenopus ribosomal gene transcription in vitro." FEBS Letters, 356(2/3): 302.
Lane N, Dean W, Erhardt S, Hajkova P, Surani A, Walter J and Reik W (2003). "Resistance of IAPs to Methylation Reprogramming May Provide a Mechanism for Epigenetic Inheritance in the Mouse." Genesis, 35(2): 88.
Lappalainen I and Vihinen M (2002). "Structural basis of ICF-causing mutations in the methyltransferase domain of DNMT3B." Protein Engineering, 15(12): 1005-1014.
Lavie L, Kitova M, Maldener E, Meese E and Mayer J (2005). "CpG Methylation Directly Regulates Transcriptional Activity of the Human Endogenous Retrovirus Family HERV-K(HML-2)." Journal of Virology, 79(2): 876-883.
Le Guezennec X, Vermeulen M, Brinkman A B, Hoeijmakers W A M, Cohen A, Lasonder E and Stunnenberg H G (2006). "MBD2/NuRD and MBD3/NuRD, Two Distinct Complexes with Different Biochemical and Functional Properties." Molecular and Cellular Biology, 26(3): 843-851.
Leahy A, Xiong J-W, Kuhnert F and Stuhlmann H (1999). "Use of developmental marker genes to define temporal and spatial patterns of differentiation during embryoid body formation." Journal of Experimental Zoology, 284(1): 67-81.
Lee J-H and Skalnik D G (2005). "CpG-binding Protein (CXXC Finger Protein 1) Is a Component of the Mammalian Set1 Histone H3-Lys4 Methyltransferase Complex, the Analogue of the Yeast Set1/COMPASS Complex." Journal of Biological Chemistry, 280(50): 41725-41731.

Lee J-H, Voo K S and Skalnik D G (2001). "Identification and Characterization of the DNA Binding Domain of CpG-binding Protein." Journal Biological Chemistry, 276(48): 44669-44676.
Lee J, Inoue K, Ono R, Ogonuki N, Kohda T, Kaneko-Ishino T, Ogura A and Ishino F (2002). "Erasing genomic imprinting memory in mouse clone embryos produced from day 11.5 primordial germ cells." Development, 129(8): 1807-1817.
Lees-Murdock D J, De Felici M and Walsh C P (2003). "Methylation dynamics of repetitive DNA elements in the mouse germ cell lineage." Genomics, 82(2): 230-237.
Leonhardt H, Page A W, Weier H U and Bestor T H (1992). "A Targeting Sequence Directs DNA Methyltransferase to Sites of DNA Replication in Mammalian Nuclei." Cell, 71(5): 865.
Li B, Zhou J, Liu P, Hu J, Jin H, Shimono Y, Takahashi M and Xu G (2007). "Polycomb protein Cbx4 promotes SUMO modification of de novo DNA methyltransferase Dnmt3a." Biochemical Journal, 405(2): 369-378.
Li E, Bestor T H and Jaenisch R (1992). "Targeted mutation of the DNA methyltransferase gene results in embryonic lethality." Cell, 69: 915-926.
Li H, Rauch T, Chen Z-X, Szabo P E, Riggs A D and Pfeifer G P (2006a). "The Histone Methyltransferase SETDB1 and the DNA Methyltransferase DNMT3A Interact Directly and Localize to Promoters Silenced in Cancer Cells." J. Biol. Chem., 281(28): 19489-19500.
Li J Y, Lees-Murdock D J, Xu G L and Walsh C P (2004). "Timing of establishment of paternal methylation imprints in the mouse." Genomics, 84(6): 952-960.
Li Q, Barkess G and Qian H (2006b). "Chromatin looping and the probability of transcription." Trends in Genetics, 22(4): 197-202.
Lin I G, Han L, Taghva A, O Brien L E and Hsieh C L (2002). "Murine De Novo Methyltransferase Dnmt3a Demonstrates Strand Asymmetry and Site Preference in the Methylation of DNA In Vitro." Molecular and Cellular Biology, 22(3): 704-723.
Ling J Q (2006). "CTCF Mediates Interchromosomal Colocalization Between Igf2/H19 and Wsb1/Nf1." Science, 312: 269-272.
Liu C L, Schreiber S L and Bernstein B E (2003). "Development and validation of a T7 based linear amplification for genomic DNA." BMC Genomics, 4(19).
Lock L F, Takagi N and Martin G R (1987). "Methylation of the Hprt gene on the inactive X occurs after chromosome inactivation." Cell, 48(1): 39-46.
Loh Y-H, Wu Q, Chew J-L, Vega V B, Zhang W*, et al.* (2006). "The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells." Nature Genetics, 38(4): 431-440.
Lorincz M C, Dickerson D R, Schmitt M and Groudine M (2004). "Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells." Nature Structural and Molecular Biology, 11(11): 1068-1075.
Lucifero D, Mertineit C, Clarke H J, Bestor T H and Trasler J M (2002). "Methylation Dynamics of Imprinted Genes in Mouse Germ Cells." Genomics, 79(4): 530-538.
Lyko F, Ramsahoye B H and Jaenisch R (2000). "Development: DNA methylation in Drosophila melanogaster." Nature, 408(6812): 538-540.

Lyko F, Ramsahoye B H, Kashevsky H, Tudor M, Mastrangelo M A, Orr-Weaver T L and Jaenisch R (1999). "Mammalian (cytosine-5) methyltransferases cause genomic DNA methylation and lethality in Drosophila." Nature Genetics, 23(3): 363-366.

Lyst M J, Nan X and Stancheva I (2006). "Regulation of MBD1-mediated transcriptional repression by SUMO and PIAS proteins." The EMBO Journal, 25: 5317-5328.

Macleod D, Charlton J, Mullins J and Bird A P (1994). "Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island." Genes and Development, 8(19): 2282.

Macleod D, Clark V H and Bird A (1999). "Absence of genome-wide changes in DNA methylation during development of the zebrafish." Nature Genetics, 23(2): 139-140.

Marahrens Y, Loring J and Jaenisch R (1998). "Role of the Xist Gene in X Chromosome Choosing." Cell, 92(5): 657-664.

Marhold J, Kramer K, Kremmer E and Lyko F (2004). "The Drosophila MBD2/3 protein mediates interactions between the MI-2 chromatin complex and CpT/A-methylated DNA." Development, 131(24): 6033-6039.

Martinowich K, Hattori D, Wu H, Fouse S, He F, Hu Y, Fan G and Sun Y E (2003). "DNA Methylation-Related Chromatin Remodeling in Activity-Dependent Bdnf Gene Regulation." Science, 302(5646): 890-893.

Matsuo K, Clay O, Takahashi T, Silke J and Schaffner W (1993). "Evidence for erosion of mouse CpG islands during mammalian evolution." Somatic Cell and Molecular Genetics, 19(6): 543-555.

Mayer-Jung C, Moras D and Timsit Y (1997). "Effect of cytosine methylation on DNA-DNA recognition at CpG steps." Journal of Molecular Biology, 270(3): 328-335.

Mayer W, Niveleau A, Walter J, Fundele R and Haaf T (2000). "Embryogenesis: Demethylation of the zygotic paternal genome." Nature, 403(6769): 501-502.

McArthur M and Thomas J O (1996). "A preference of histone H1 for methylated DNA." Embo Journal, 15(7): 1705-1714.

McClelland M and Ivarie R (1982). "Asymmetrical distribution of CpG in an 'average' mammalian gene." Nucleic Acids Research, 10(23): 7865–7877.

McCool K W, Xu X, Singer D B, Murdoch F E and Fritsch M K (2007). "The Role of Histone Acetylation in Regulating Early Gene Expression Patterns during Early Embryonic Stem Cell Differentiation." Journal of Biological Chemistry, 282: 6696.

Meehan R R, Lewis J D, McKay S, Kleiner E L and Bird A P (1989). "Identification of a mammalian protein that binds specifically to DNA containing methylated CpGs." Cell, 58: 499-507.

Mertineit C, Yoder J A, Taketo T, Laird D W, Trasler J M and Bestor T H (1998). "Sex-specific exons control DNA methyltransferase in mammalian germ cells." Development, 125(5): 889-897.

Mette M F, Aufsatz W, van der Winden J, Matzke M A and Matzke A J M (2000). "Transcriptional silencing and promoter methylation triggered by double-stranded RNA." Embo Journal, 19(19): 5194-5201.

Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, Rabkin S, Guo N, Muruganujan A, Doremieux O, Campbell M J, Kitano H and Thomas P D

(2005). "The PANTHER database of protein families, subfamilies, functions and pathways." Nucleic Acids Research, 33(suppl_1): D284-288.
Millar C B, Guy J, Sansom O J, Selfridge J, MacDougall E, Hendrich B, Keightley P D, Bishop S M, Clarke A R and Bird A (2002). "Enhanced CpG Mutability and Tumorigenesis in MBD4-Deficient Mice." Science, 297(5580): 403-405.
Milne T A, Dou Y, Martin M E, Brock H W, Roeder R G and Hess J L (2005). "From The Cover: MLL associates specifically with a subset of transcriptionally active target genes." Proceedings- National Academy of Sciences USA, 102(41): 14765-14770.
Minucci S, Botquin V, Yeom Y, Dey A, Sylvester I, Zand D J, Ohbo K, Ozato K and Schoeler H R (1996). "Retinoic acid-mediated down-regulation of Oct3/4 coincides with the loss of promoter occupancy in vivo." Embo Journal, 15(4): 888-899.
Mitchell A R, Jeppesen P, Nicol L, Morrison H and Kipling D (1996). "Epigenetic control of mammalian centromere protein binding: does DNA methylation have a role?" Journal of Cell Science, 109(9): 2199-2206.
Moore T and Haig D (1991). "Genomic imprinting in mammalian development: a parental tug-of-war." Trends in Genetics, 7(2): 45-49.
Morgan H D, Dean W, Coker H A, Reik W and Petersen-Mahrt S K (2004). "Activation-induced Cytidine Deaminase Deaminates 5-Methylcytosine in DNA and Is Expressed in Pluripotent Tissues: IMPLICATIONS FOR EPIGENETIC REPROGRAMMING." Journal Biological Chemistry, 279(50): 52353-52360.
Mouse Genome Sequencing Consortium, (2002). "Initial sequencing and comparative analysis of the mouse genome" Nature, 420: 520-562.
Muiznieks I and Doerfler W (1994). "The topology of the promoter of RNA polymerase II- and III-transcribed genes is modified by the methylation of 5'-CG-3' dinucleotides." Nucleic Acids Research, 22(13): 2568.
Muller J, Hart C M, Francis N J, Vargas M L, Sengupta A, Wild B, Miller E L, O'Connor M B, Kingston R E and Simon J A (2002). "Histone Methyltransferase Activity of a Drosophila Polycomb Group Repressor Complex." Cell, 111(2): 197-208.
Mummaneni P, Yates P, Simpson J, Rose J and Turker M S (1998). "The primary function of a redundant Sp1 binding site in the mouse aprt gene promoter is to block epigenetic gene inactivation." Nucleic Acids Research, 26(22): 5163-5169.
Murrell A, Heeson S and Reik W (2004). "Interaction between differentially methylated regions partitions the imprinted genes Igf2 and H19 into parent-specific chromatin loops." Nature Genetics, 36: 889-893.
Mutskov V and Felsenfeld G (2004). "Silencing of transgene transcription precedes methylation of promoter DNA and histone H3 lysine 9." Embo Journal, 23(1): 138-149.
Nakamura T, Arai Y, Umehara H, Masuhara M, Kimura T, Taniguchi H, Sekimoto T, Ikawa M, Yoneda Y, Okabe M, Tanaka S, Shiota K and Nakano T (2007). "PGC7/Stella protects against DNA demethylation in early embryogenesis." Nature Cell Biology, 9(1): 64-71.
Nan X, Ng H-H, Johnson C A, Laherty C D, Turner B M, Eisenman R N and Bird A (1998). "Transcriptional repression by the methyl-CpG-binding protein

MeCP2 involves a histone deacetylase complex." Nature, 393(6683): 386-389.
Naveh-Many T and Cedar H (1982). "Topographical distribution of 5-methylcytosine in animal and plant DNA." Molecular and Cellular Biology, 2(7): 758-762.
Ng H-H, Zhang Y, Hendrich B, Johnson C A, Turner B M, Erdjument-Bromage H, Tempst P, Reinberg D and Bird A (1999). "MBD2 is a transcriptional repressor belonging to the MeCP1 histone deacetylase complex." Nature Genetics, 23(1): 58-61.
Ng H H, Robert F, Young R A and Struhl K (2003). "Targeted Recruitment of Set1 Histone Methylase by Elongating Pol II Provides a Localized Mark and Memory of Recent Transcriptional Activity." Molecular Cell, 11(3): 709-720.
Nichols J, Zevnik B, Anastassiadis K, Niwa H, Klewe-Nebenius D, Chambers I, Schoeler H and Smith A (1998). "Formation of Pluripotent Stem Cells in the Mammalian Embryo Depends on the POU Transcription Factor Oct4." Cell, 95(3): 379-391.
Nightingale K and Wolffe A P (1995). "Methylation at CpG sequences does not influence histone H1 binding to a nucleosome including a Xenopus borealis 5 S rRNA gene." Journal of Biological Chemistry, 270(9): 4197.
Nishimoto M, Fukushima A, Okuda A and Muramatsu M (1999). "The Gene for the Embryonic Stem Cell Coactivator UTF1 Carries a Regulatory Element Which Selectively Interacts with a Complex Composed of Oct-3/4 and Sox-2." Molecular and Cellular Biology, 19(8): 5453-5465.
Niwa H, Miyazaki J and Smith A G (2000). "Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells." Nature Genetics, 24(4): 372-376.
Nordhoff V, Hubner K, Bauer A, Orlova I, Malapetsa A and Scholer H R (2001). "Comparative analysis of human, bovine, and murine Oct-4 upstream promoter sequences." Mammalian Genome, 12(4): 309-317.
O'Neill M J (2005). "The influence of non-coding RNAs on allele-specific gene expression in mammals." Human Molecular Genetics, 14(suppl_1): R113-120.
Oakes C C, La Salle S, Smiraglia D J, Robaire B and Trasler J M (2007a). "Developmental acquisition of genome-wide DNA methylation occurs prior to meiosis in male germ cells." Developmental Biology, 307(2): 368-379.
Oakes C C, La Salle S, Smiraglia D J, Robaire B and Trasler J M (2007b). "A unique configuration of genome-wide DNA methylation patterns in the testis." Proceedings- National Academy of Sciences USA, 104: 228-233.
Obata Y and Kono T (2002). "Maternal primary imprinting is established at a specific time for each gene throughout oocyte growth." Journal of Biological Chemistry, 277(7): 5285-5289.
Ogawa Y and Lee J T (2003). "Xite, X-Inactivation Intergenic Transcription Elements that Regulate the Probability of Choice." Molecular Cell, 11(3): 731-743.
Ohki I, Shimotake N, Fujita N, Nakao M and Shirakawa M (1999). "Solution structure of the methyl-CpG-binding domain of the methylation-dependent transcriptional repressor MBD1." The EMBO Journal, 18: 6653-6661.

Ohlsson R, Renkawitz R and Lobanenkov V (2001). "CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease." Trends in Genetics, 17(9): 520-527.

Okano M, Bell D W, Haber D A and Li E (1999). "DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development." Cell, 99(3): 247-257.

Okazawa H, Okamoto K, Ishino F, Ishino-Kaneko T, Takeda S, Toyoda Y, Muramatsu M and Hamada H (1991). "The oct3 gene, a gene for an embryonic transcription factor, is controlled by a retinoic acid repressible enhancer." The EMBO Journal, 10(10): 2997–3005.

Okita K, Ichisaka T and Yamanaka S (2007). "Generation of germline-competent induced pluripotent stem cells." Nature, 448(7151): 313-317.

Okumura-Nakanishi S, Saito M, Niwa H and Ishikawa F (2005). "Oct-3/4 and Sox2 Regulate Oct-3/4 Gene in Embryonic Stem Cells." Journal of Biological Chemistry, 280: 5307-5317.

Ooi S K T, Qiu C, Bernstein E, Li K, Jia D, Yang Z, Erdjument-Bromage H, Tempst P, Lin S-P, Allis C D, Cheng X and Bestor T H (2007). "DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA." Nature, 448(7154): 714-717.

Oswald J, Engemann S, Lane N, Mayer W, Olek A, Fundele R, Dean W, Reik W and Walter J (2000). "Active demethylation of the paternal genome in the mouse zygote." Current Biology, 10(8): 475-478.

Ovitt C E and Schoeler H R (1998). "The molecular biology of Oct-4 in the early mouse embryo." Molecular Human Reproduction, 4(11): 1021-1031.

Papait R, Pistore C, Negri D, Pecoraro D, Cantarini L and Bonapace I M (2007). "Np95 Is Implicated in Pericentromeric Heterochromatin Replication and in Major Satellite Silencing." Molecular and Cellular Biology, 18(3): 1098-1106.

Park J-i, Kim S W, Lyons J P, Ji H, Nguyen T T, Cho K, Barton M C, Deroo T, Vleminckx K and McCrea P D (2005). "Kaiso/p120-Catenin and TCF/[beta]-Catenin Complexes Coordinately Regulate Canonical Wnt Gene Targets." Developmental Cell, 8(6): 843-854.

Penny G D, Kay G F, Sheardown S A, Rastan S and Brockdorff N (1996). "Requirement for Xist in X chromosome inactivation." Nature, 379(6561): 131-137.

Pesole G, Bernardi G and Saccone C (1999). "Isochore specificity of AUG initiator context of human genes." FEBS Letters, 464(1-2): 60-62.

Petronzelli F, Riccio A, Markham G D, Seeholzer S H, Stoerker J, Genuardi M, Yeung A T, Matsumoto Y and Bellacosa A (2000). "Biphasic Kinetics of the Human DNA Repair Protein MED1 (MBD4), a Mismatch-specific DNA N-Glycosylase." Journal Biological Chemistry, 275(42): 32422-32429.

Pikarsky E, Sharir H, Ben-Shushan E and Bergman Y (1994). "Retinoic Acid Represses Oct-3/4 Gene Expression through Several Retinoic Acid-Responsive Elements Located in the Promoter-Enhancer Region." Molecular and Cellular Biology, 14(2): 1026.

Pollack Y, Kogan N and Golenser J (1991). "Plasmodium falciparum: Evidence for a DNA methylation pattern." Experimental Parasitology, 72(4): 339-344.

Ponger L, Duret L and Mouchiroud D (2001). "Determinants of CpG Islands: Expression in Early Embryo and Isochore Structure." Genome Research, 11(11): 1854-1860.

Prokhortchouk A, Hendrich B, Jorgensen H, Ruzov A, Wilm M, Georgiev G, Bird A and Prokhortchouk E (2001). "The p120 catenin partner Kaiso is a DNA methylation-dependent transcriptional repressor." Genes and Development, 15(13): 1613-1618.

Prokhortchouk A, Sansom O, Selfridge J, Caballero I M, Salozhin S*, et al.* (2006). "Kaiso-Deficient Mice Show Resistance to Intestinal Cancer." Molecular and Cellular Biology, 26(1): 199-208.

Ramsahoye B H, Biniszkiewicz D, Lyko F, Clark V, Bird A P and Jaenisch R (2000). "Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a." Proceedings- National Academy of Sciences USA, 97(10): 5237-5242.

Regev A, Lamb M J and Jablonka E (1998). "The Role of DNA Methylation in Invertebrates: Developmental Regulation or Genome Defense?" Molecular Biology and Evolution, 15(7): 880-891.

Reik W and Walter J (2001). "Genomic imprinting: parental influence on the genome." Nature Reviews Genetics, 2(1): 21-32.

Rhee I, Bachman K E, Park B H, Jair K W, Yen R W C, Schuebel K E, Cui H, Feinberg A P, Lengauer C and Kinzler K W (2002). "DNMT1 and DNMT3b cooperate to silence genes in human cancer cells." Nature(6880): 552-555.

Riggs A D and Pfeifer G P (1992). "X-chromosome inactivation and cell memory." Trends in Genetics, 8(5): 169-174.

Robertson K D, Ait-Si-Ali S, Yokochi T, Wade P A, Jones P L and Wolffe A P (2000). "DNMT1 forms a complex with Rb, E2F1 and HDAC1 and represses transcription from E2F-responsive promoters." Nature Genetics, 25(3): 338-342.

Robertson K D, Uzvolgyi E, Liang G, Talmadge C, Sumegi J, Gonzales F A and Jones P A (1999). "The human DNA methyltransferases (DNMTs) 1, 3a and 3b: coordinate mRNA expression in normal tissues and overexpression in tumors." Nucleic Acids Research, 27(11): 2291-2298.

Robinson P N, Bohme U, Lopez R, Mundlos S and Nurnberg P (2004). "Gene-Ontology analysis reveals association of tissue-specific 5prime CpG-island genes with development and embryogenesis." Human Molecular Genetics, 13(17): 1969-1978.

Rodda D J, Chew J L, Lim L H, Loh Y H, Wang B, Ng H H and Robson P (2005). "Transcriptional Regulation of Nanog by OCT4 and SOX2." Journal of Biological Chemistry, 280: 24731-24737.

Roloff T C, Ropers H H and Nuber U A (2003). "Comparative study of methyl-CpG-binding domain proteins." BMC Genomics, 4.

Rountree M R, Bachman K E and Baylin S B (2000). "DNMT1 binds HDAC2 and a new co-repressor, DMAP1, to form a complex at replication foci." Nature Genetics, 25(3): 269-278.

Ruzov A, Dunican D S, Prokhortchouk A, Pennings S, Stancheva I, Prokhortchouk E and Meehan R R (2004). "Kaiso is a genome-wide repressor of transcription that is essential for amphibian development." Development, 131(24): 6185-6194.

Sansom O J, Berger J, Bishop S M, Hendrich B, Bird A and Clarke A R (2003). "Deficiency of Mbd2 suppresses intestinal tumorigenesis." Nature Genetics, 34(2): 145-147.
Santos-Rosa H, Schneider R, Bannister A J, Sherriff J, Bernstein B E, Emre N C T, Schreiber S L, Mellor J and Kouzarides T (2002). "Active genes are tri-methylated at K4 of histone H3." Nature(6905): 407-410.
Sarg B, Helliger W, Talasz H, Koutzamani E and Lindner H H (2004). "Histone H4 Hyperacetylation Precludes Histone H4 Lysine 20 Trimethylation." Journal of Biological Chemistry, 279: 53458-53464.
Sarraf S A and Stancheva I (2004). "Methyl-CpG Binding Protein MBD1 Couples Histone H3 Methylation at Lysine 9 by SETDB1 to DNA Replication and Chromatin Assembly." Molecular Cell, 15(4): 595-605.
Sato N, Kondo M and Arai K i (2006). "The orphan nuclear receptor GCNF recruits DNA methyltransferase for Oct-3/4 silencing." Biochemical and Biophysical Research Communications, 344(3): 845-851.
Saxonov S, Berg P and Brutlag D L (2006). "A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters." Proceedings- National Academy of Sciences USA, 103: 1412.
Schöler H R, Balling R, Hatzopoulos A K, Suzuki N and Gruss P (1989a). "Octamer binding proteins confer transcriptional activity in early mouse embryogenesis." The EMBO Journal, 8(9): 2551-2557.
Schöler H R, Hatzopoulos A K, Balling R, Suzuki N and Gruss P (1989b). "A family of octamer-specific proteins present during mouse embryogenesis: evidence for germline-specific expression of an Oct factor." The EMBO Journal, 8(9): 2543-50.
Schoorlemmer J, van Puijenbroek A, van Den Eijnden M, Jonk L, Pals C and Kruijer W (1994). "Characterization of a negative retinoic acid response element in the murine Oct4 promoter." Mol. Cell. Biol., 14(2): 1122-1136.
Seki Y, Hayashi K, Itoh K, Mizugaki M, Saitou M and Matsui Y (2005). "Extensive and orderly reprogramming of genome-wide chromatin modifications associated with specification and early development of germ cells in mice." Developmental Biology, 278(2): 440-458.
Selker E U (1990). "Premeiotic Instability of Repeated Sequences in Neurospora Crassa." Annual Review of Genetics, 24(1): 579-613.
Selker E U, Tountas N A, Cross S H, Margolin B S, Murphy J G, Bird A P and Freitag M (2003). "The methylated component of the Neurospora crassa genome." Nature, 422(6934): 893-897.
Shabalina S A, Ogurtsov A Y, Lipman D J and Kondrashov A S (2003). "Patterns in interspecies similarity correlate with nucleotide composition in mammalian 3'UTRs." Nucleic Acids Research, 31(18): 5433-5439.
Shapiro J A and Von Sternberg R (2005). "Why repetitive DNA is essential to genome function." Biological Reviews- Cambridge Philosophical Society, 80: 227-250.
Shimizu T S, Takahashi K and Tomita M (1997). "CpG distribution patterns in methylated and non-methylated species." Gene, 205(1/2): 103-107.
Shin Voo K, Carlone D L, Jacobsen B M, Flodin A and Skalnik D G (2000). "Cloning of a Mammalian Transcriptional Activator That Binds Unmethylated CpG Motifs and Shares a CXXC Domain with DNA

Methyltransferase, Human Trithorax, and Methyl-CpG Binding Domain Protein 1." Molecular and Cellular Biology, 20(6): 2108-2121.
Shiota K, Kogo Y, Ohgane J, Imamura T, Urano A, Nichino K, Tanaka S and Hattori N (2002). "Epigenetic marks by DNA methylation specific to stem, germ and somatic cells in mice." Genes to Cells, 7: 961-970.
Shovlin T C, Bourc'his D, La Salle S, O'Doherty A, Trasler J M, Bestor T H and Walsh C P (2007). "Sex-specific promoters regulate Dnmt3L expression in mouse germ cells." Human Reproduction, 22(2): 457-467.
Simmen M W, Leitgeb S, Charlton J, Jones S J N, Harris B R, Clark V H and Bird A (1999). "Nonmethylated Transposable Elements and Methylated Genes in a Chordate Genome." Science: 1164-1167.
Sleutels F, Tjon G, Ludwig T and Barlow D P (2003). "Imprinted silencing of Slc22a2 and Slc22a3 does not need transcriptional overlap between Igf2r and Air." The EMBO Journal, 22(14): 3696–3704.
Sleutels F, Zwart R and Barlow D P (2002). "The non-coding Air RNA is required for silencing autosomal imprinted genes." Nature, 415(6873): 810-813.
Smyth G K (2004). "Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments." Statistical Applications in Genetics and Molecular Biology, 3(1).
Song F, Smith J F, Kimura M T, Morrow A D, Matsuyama T, Nagase H and Held W A (2005). "Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression." Proceedings of the National Academy of Sciences, 102(9): 3336-3341.
Spada F, Haemmer A, Kuch D, Rothbauer U, Schermelleh L, Kremmer E, Carell T, Langst G and Leonhardt H (2007). "DNMT1 but not its interaction with the replication machinery is required for maintenance of DNA methylation in human cells." Journal of Cell Biology, 176: 565-572.
Spencer V A and Davie J R (1999). "Role of covalent modifications of histones in regulating gene expression." Gene, 240(1): 1-12.
Stancheva I, Collins A L, Van den Veyver I B, Zoghbi H and Meehan R R (2003). "A Mutant Form of MeCP2 Protein Associated with Human Rett Syndrome Cannot Be Displaced from Methylated DNA by Notch in Xenopus Embryos." Molecular Cell, 12(2): 425-435.
Stancheva I and Meehan R R (2000). "Transient depletion of xDnmt1 leads to premature gene activation in Xenopus embryos." Genes and Development, 14(3): 313-327.
Stavropoulos N, Lu N and Lee J T (2001). "A functional role for Tsix transcription in blocking Xist RNA accumulation but not in X-chromosome choice." Proceedings- National Academy of Sciences USA, 98(18): 10232-10237.
Storey J D and Tibshirani R (2003). "Statistical significance for genomewide studies." Proceedings- National Academy of Sciences USA, 100(16): 9440.
Su A I, Cooke M P, Ching K A, Hakak Y, Walker J R, Wiltshire T, Orth A P, Vega R G, Sapinoso L M, Moqrich A, Patapoutian A, Hampton G M, Schultz P G and Hogenesch J B (2002). "Large-scale analysis of the human and mouse transcriptomes." Proceedings of the National Academy of Sciences: 012025199.

Suetake I, Miyazaki J, Murakami C, Takeshima H and Tajima S (2003). "Distinct Enzymatic Properties of Recombinant Mouse DNA Methyltransferases Dnmt3a and Dnmt3b." Journal of Biochemistry, 133(6): 737-744.

Sugimoto M and Abe K (2007). "X chromosome reactivation initiates in nascent primordial germ cells in mice." PLoS Genetics, 3(7): e116.

Suzuki M M, Kerr A R W, De Sousa D and Bird A (2007). "CpG methylation is targeted to transcription units in an invertebrate genome." Genome Research, 17: 625-631.

Symer D E, Connelly C, Szak S T, Caputo E M, Cost G J, Parmigiani G and Boeke J D (2002). "Human L1 Retrotransposition Is Associated with Genetic Instability In Vivo." Cell, 110(3): 327-338.

Tachibana M, Sugimoto K, Nozaki M, Ueda J, Ohta T, Ohki M, Fukuda M, Takeda N, Niida H, Kato H and Shinkai Y (2002). "G9a histone methyltransferase plays a dominant role in euchromatic histone H3 lysine 9 methylation and is essential for early embryogenesis." Genes and Development, 16(14): 1779-1791.

Takahashi K and Yamanaka S (2006). "Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors." Cell, 126(4): 663-676.

Takai D and Jones P A (2002). "Comprehensive analysis of CpG islands in human chromosomes 21 and 22." Proceedings- National Academy of Sciences USA, 99(6): 3740-3745.

Takeshima H, Suetake I, Shimahara H, Ura K, Tate S-i and Tajima S (2006). "Distinct DNA Methylation Activity of Dnmt3a and Dnmt3b towards Naked and Nucleosomal DNA." Journal of Biochem, 139(3): 503-515.

Tan C P and Nakielny S (2006). "Control of the DNA Methylation System Component MBD2 by Protein Arginine Methylation." Molecular and Cellular Biology, 26(19): 7224–7235.

Terranova R, Agherbi H, Boned A, Meresse S and Djabali M (2006). "Histone and DNA methylation defects at Hox genes in mice expressing a SET domain-truncated form of Mll." Proceedings- National Academy of Sciences USA, 103(17): 6629-6634.

Thomas P D, Campbell M J, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A and Narechania A (2003). "PANTHER: A Library of Protein Families and Subfamilies Indexed by Function." Genome Research, 13(9): 2129-2141.

Thorvaldsen J, Duran K and Bartolomei M (1998a). "The H19 differentially-methylated region is required for imprinted expression of H19 and Igf2." Abstracts of Papers Presented at the Meeting on Mouse Molecular Genetics: 270.

Thorvaldsen J L, Duran K L and Bartolomei M S (1998b). "Deletion of the H19 differentially methylated domain results in loss of imprinted expression of H19 and Igf2." Genes and Development, 12(23): 3693-3702.

Ting A H, McGarvey K M and Baylin S B (2006). "The cancer epigenome-components and functional correlates." Genes and Development, 20: 3215-3231.

Tran R K, Henikoff J G, Zilberman D, Ditt R F, Jacobsen S E and Henikoff S (2005). "DNA Methylation Profiling Identifies CG Methylation Clusters in Arabidopsis Genes." Current Biology, 15(2): 154-159.

Tsumura A, Hayakawa T, Kumaki Y, Takebayashi S i, Sakaue M, Matsuoka C, Shimotohno K, Ishikawa F, Li E and Ueda H R (2006). "Maintenance of self-renewal ability of mouse embryonic stem cells in the absence of DNA methyltransferases Dnmt1, Dnmt3a and Dnmt3b." Genes to Cells, 11(7): 805-814.

Tufarelli C, Stanley J A S, Garrick D, Sharpe J A, Ayyub H, Wood W G and Higgs D R (2003). "Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease." Nature Genetics, 34(2): 157-165.

Turek-Plewa J and Jagodzinski P P (2005). "The Role of Mammalian DNA Methyltransferases in the Regulation of Gene Expression." Cellular and Molecular Biology Letters, 10(4): 631-648.

Turker M S (2002). "Gene silencing in mammalian cells and the spread of DNA methylation." Oncogene, 21(35): 5388-5393.

Tweedie S, Charlton J, Clark V and Bird A (1997). "Methylation of genomes and genes at the invertebrate-vertebrate boundary." Molecular and Cellular Biology, 17(3): 1469-1475.

Ueda Y, Okano M, Williams C, Chen T, Georgopoulos K and Li E (2006). "Roles for Dnmt3b in mammalian development: a mouse model for the ICF syndrome." Development, 133(6): 1183-1192.

Unoki M and Nakamura Y (2003). "Methylation at CpG islands in intron 1 of EGR2 confers enhancer-like activity." FEBS Letters, 554(1-2): 67-72.

Unoki M, Nishidate T and Nakamura Y (2003). "ICBP90, an E2F-1 target, recruits HDAC1 and binds to methyl-CpG through its SRA domain." Oncogene, 23(46): 7601-7610.

Vakoc C R, Mandat S A, Olenschock B A and Blobel G A (2005). "Histone H3 Lysine 9 Methylation and HP1gamma Are Associated with Transcription Elongation through Mammalian Chromatin. (Poster # 892-I)." Blood, 106(1): 1734.

Van den Veyver I B and Zoghbi H Y (2001). "Mutations in the gene encoding methyl-CpG-binding protein 2 cause Rett syndrome." Brain and Development, 23(Supplement 1): S147-S151.

Varriale A and Bernardi G (2006a). "DNA methylation and body temperature in fishes." Gene, 385: 111-121.

Varriale A and Bernardi G (2006b). "DNA methylation in reptiles." Gene, 385: 122-127.

Viegas-Pequignot E, Dutrillaux B and Thomas G (1988). "Inactive X Chromosome Has the Highest Concentration of Unmethylated Hha I Sites." Proceedings of the National Academy of Sciences, 85(20): 7657-7660.

Villa R, Morey L, Raker V A, Buschbeck M, Gutierrez A, De Santis F, Corsaro M, Varas F, Bossi D and Minucci S (2006). "The methyl-CpG binding protein MBD1 is required for PML-RARalpha function." Proceedings- National Academy of Sciences USA, 103: 1400.

Vinogradov A E (2003). "Isochores and tissue-specificity." Nucleic Acids Research, 31(17): 5212-5220.

Vire E, Brenner C, Deplus R, Blanchon L, Fraga M*, et al.* (2006). "The Polycomb group protein EZH2 directly controls DNA methylation." Nature, 439(7078): 871-874.
Wade P A, Gegonne A, Jones P L, Ballestar E, Aubry F and Wolffe A P (1999). "Mi-2 complex couples DNA methylation to chromatin remodelling and histone deacetylation." Nature Genetics, 23(1): 62-66.
Wakefield R I D, Smith B O, Nan X, Free A, Soteriou A, Uhrin D, Bird A P and Barlow P N (1999). "The solution structure of the domain from MeCP2 that binds to methylated DNA." Journal of Molecular Biology, 291(5): 1055-1065.
Walsh C P, Chaillet J R and Bestor T H (1998). "Transcription of IAP endogenous retroviruses is constrained by cytosine methylation." Nat Genet, 20(2): 116-117.
Wang R Y H, Gehrke C W and Ehrlich M (1980). "Comparison of bisulfite modification of 5-methyldeoxycytidine and deoxycytidine residues." Nucl. Acids Res., 8(20): 4777-4790.
Wang Y (2006). "Functional CpG Methylation System in a Social Insect." Science, 314: 645-646.
Wang Y, Lu J, Lee R and Clarke R (2002). "Iterative normalization of cDNA microarray data." IEEE Transactions on Information Technology in Biomedicine, 6(1): 29-37.
Wassenegger M, Heimes S, Riedel L and Sänger H L (1994). "RNA-directed de novo methylation of genomic sequences in plants." Cell, 76: 567-576.
Wassenegger M and Krczal G (2006). "Nomenclature and functions of RNA-directed RNA polymerases." Trends in Plant Science, 11(3): 142-151.
Watanabe D, Suetake I, Tada T and Tajima S (2002). "Stage- and cell-specific expression of Dnmt3a and Dnmt3b during embryogenesis." Mechanisms of Development, 118(1-2): 187-190.
Watson J D and Crick F H C (1953). "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid." Nature, 171: 737-738.
Weber M, Davies J J, Wittig D, Oakeley E J, Haase M, Lam W L and Schubeler D (2005). "Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells." Nature Genetics, 37(8): 853-862.
Weber M, Hellmann I, Stadler M B, Ramos L, Paabo S, Rebhan M and Schubeler D (2007). "Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome." Nature Genetics, 39(4): 457-466.
Weisenberger D J, Velicescu M, Preciado-Lopez M A, Gonzales F A, Tsai Y C, Liang G and Jones P A (2002). "Identification and characterization of alternatively spliced variants of DNA methyltransferase 3a in mammalian cells." Gene, 298(1): 91-99.
Wernig M, Meissner A, Foreman R, Brambrink T, Ku M, Hochedlinger K, Bernstein B E and Jaenisch R (2007). "In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state." Nature, 448(7151): 318-324.
White G P, Watt P M, Holt B J and Holt P G (2002). "Differential Patterns of Methylation of the IFN-gamma Promoter at CpG and Non-CpG Sites Underlie Differences in IFN-gamma Gene Expression Between Human

Neonatal and Adult CD45RO^- T Cells." Journal of Immunology, 168(6): 2820-2827.
Woo H R, Pontes O, Pikaard C S and Richards E J (2007). "VIM1, a methylcytosine-binding protein required for centromeric heterochromatinization." Genes and Development, 21(3): 267-277.
Wu D Y and Yao Z (2006). "Functional analysis of two Sp1/Sp3 binding sites in murine Nanog gene promoter." Cell Res, 16(3): 319-322.
Wu P, Qiu C, Sohail A, Zhang X, Bhagwat A S and Cheng X (2003). "Mismatch Repair in Methylated DNA.Structure and activity of the mismatch-specific thymine glycosylase domain of methyl-CpG-binding protein MBD4." Journal Biological Chemistry, 278(7): 5285-5291.
Wu Q, Chen X, Zhang J, Loh Y-H, Low T-Y, Zhang W, Zhang W, Sze S-K, Lim B and Ng H-H (2006). "Sall4 Interacts with Nanog and Co-occupies Nanog Genomic Sites in Embryonic Stem Cells." Journal of Biological Chemistry, 281(34): 24090-24094.
Wutz A and Jaenisch R (2000). "A Shift from Reversible to Irreversible X Inactivation Is Triggered during ES Cell Differentiation." Molecular Cell, 5(4): 695-705.
Wutz A, Smrzka O W, Schweifer N, Schellander K, Wagner E F and Barlow D P (1997). "Imprinted expression of the Igf2r gene depends on an intronic CpG island." Nature, 389(6652): 745-749.
Xiao T, Hall H, Kizer K O, Shibata Y, Hall M C, Borchers C H and Strahl B D (2003). "Phosphorylation of RNA polymerase II CTD regulates H3 methylation in yeast." Genes and Development, 17(5): 654-663.
Xu N, Tsai C L and Lee J T (2006). "Transient Homologous Chromosome Pairing Marks the Onset of X Inactivation." Science, 311: 1149-1152.
Yagi T, Tokunaga T, Furuta Y, Nada S, Yoshida M, Tsukada T, Saga Y, Takeda N, Ikawa Y and Aizawa S (1993). "A novel ES cell line, TT2, with high germline-differentiating potency." Analytical Biochemistry, 214(1): 70-76.
Yamashita R, Suzuki Y, Sugano S and Nakai K (2005). "Genome-wide analysis reveals strong correlation between CpG islands with nearby transcription start sites of genes and their tissue specificity." Gene, 350(2): 129-136.
Yeo S, Jeong S, Kim J, Han J S, Han Y M and Kang Y K (2007). "Characterization of DNA methylation change in stem cell marker genes during differentiation of human embryonic stem cells." Biochemical and Biophysical Research Communications, 359(3): 536-542.
Yeom Y I, Fuhrmann G, Ovitt C E, Brehm A, Ohbo K, Gross M, Huebner K and Schoeler H R (1996). "Germline regulatory element of Oct-4 specific for the totipotent cycle of embryonal cells." Development, 122(3): 881-894.
Yoder J A, Walsh C P and Bestor T H (1997). "Cytosine methylation and the ecology of intragenomic parasites." Trends in Genetics, 13(8): 335-340.
Yoon B, Herman H, Hu B, Park Y J, Lindroth A, Bell A, West A G, Chang Y, Stablewski A, Piel J C, Loukinov D I, Lobanenkov V V and Soloway P D (2005). "Rasgrf1 Imprinting Is Regulated by a CTCF-Dependent Methylation-Sensitive Enhancer Blocker." Molecular and Cellular Biology, 25(24): 11184-11190.

Yoon H-G, Chan D W, Reynolds A B, Qin J and Wong J (2003). "N-CoR Mediates DNA Methylation-Dependent Repression through a Methyl CpG Binding Protein Kaiso." Molecular Cell, 12(3): 723-734.

Yuan H, Corbi N, Basilico C and Dailey L (1995). "Developmental-specific activity of the FGF-4 enhancer requires the synergistic action of Sox2 and Oct-3." Genes and Development, 9(21): 2635.

Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan S W, Chen H, Henderson I R, Shinn P, Pellegrini M and Jacobsen S E (2006). "Genome-wide High-Resolution Mapping and Functional Analysis of DNA Methylation in Arabidopsis." Cell, 126(6): 1189-1201.

Zhang Y, Ng H-H, Erdjument-Bromage H, Tempst P, Bird A and Reinberg D (1999). "Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation." Genes and Development, 13(15): 1924-1935.

Zhao X, Ueba T, Christie B R, Barkho B, McConnell M J, Nakashima K, Lein E S, Eadie B D, Willhoite A R, Muotri A R, Summers R G, Chun J, Lee K-F and Gage F H (2003). "Mice lacking methyl-CpG binding protein 1 have deficits in adult neurogenesis and hippocampal function." Proceedings- National Academy of Sciences USA, 100(11): 6777-6782.

Zilberman D, Gehring M, Tran R K, Ballinger T and Henikoff S (2007). "Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription." Nature Genetics, 39(1): 61-69.

Zwart R, Sleutels F, Wutz A, Schinkel A H and Barlow D P (2001). "Bidirectional action of the Igf2r imprint control element on upstream and downstream imprinted genes." Genes and Development, 15(18): 2361-2366.